Methods for Global Characterization of Chromatin Regulators in Human Cells

Dissertation Advisor: Professor Bradley Bernstein Vicky Weijie Zhou

Methods for Global Characterization of Chromatin Regulators in Human Cells

Abstract

Chromatin is a multi-layered structure composed of DNA, nucleosomes, histone modifications, and associated proteins that critically affects genome function. Recently developed sequencing technologies enable genomewide characterization of certain aspects of chromatin structure, including nucleosome positioning and histone modifications. However, chromatin proteins present several challenges due to their dynamic nature and variable association with DNA. Chromatin proteins such as Polycomb regulators and heterochromatic factors play critical and global roles in epigenetic repression and hence new approaches are needed for their study.

We first sought to identify sequences that recruit Polycomb repressive complex 2 (PRC2) in mammalian cells. We combined chromatin immunoprecipitation with sequencing (ChIP-seq) to map the candidate transcription factor YY1, and found that it does not correlate with PRC2 localization, suggesting that YY1 is not directly involved in PRC2 recruitment. We also identified GC-rich sequences that are necessary and sufficient for PRC2 recruitment. Yet attempts to map additional Polycomb proteins and other repressors using ChIP-seq proved difficult.

Since chromatin proteins are often broadly, secondarily or transiently bound to DNA, they are difficult to crosslink. Antibody quality also varies, further hampering ChIP-seq technology. Here, we adapt DamID, a method for mapping chromatin regulators that uses a

fusion enzyme and that does not rely on crosslinking or antibodies, for high-throughput sequencing. We show that DamID-seq can be used to globally characterize chromatin repressors in human cells.

We used DamID-seq to map the binding of 12 chromodomain-containing and related proteins in K562 cells. We found that these proteins cluster into two modules: 1) Polycomb-related and 2) heterochromatin-related. Polycomb proteins bind developmental genes, while heterochromatin proteins bind broad olfactory receptor (OR) and zinc finger (ZNF) domains. Surprisingly, unlike other Polycomb proteins, CBX2 uniquely binds genes involved with modifying proteins.

Our findings advance the model that the genome is compartmentalized into domains, and identify the distinct protein components that associate respectively with Polycomb and heterochromatin domains in human cells. We expect that DamID-seq, along with further advancements in characterizing the three-dimensional organization of chromatin, will bring us towards a better understanding of the role of chromatin in differentiation, development, and disease.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ACKNOWLEDGMENTS

I am deeply indebted to countless individuals during my past five years of graduate studies.

First and foremost, I would like to express my gratitude to my thesis advisor, Prof. Bradley Bernstein, for his unwavering support and guidance on my projects, immediate feedback on my progress, and keen attention to my professional development. This Thesis would not have been possible without the critical opportunities he provided and the engaging environment he created.

I would also like to thank numerous members of the Bernstein Lab: Dr. Yoshiko Mito and Daniel Fernandez for a fruitful collaboration on DamID-seq technology; Dr. Eric Mendenhall for advice and collaboration on the BAC project; Dr. Alon Goren for advice and collaboration on the review paper; Dr. Andrew Chi for mentorship during my rotation project; Dr. Esther Rheinbay, Richard Koche, and Dr. Jiang Zhu for teaching me computational analysis; Dr. Manching Ku for general and technical advice; Dr. Oren Ram for immediate help on DamID-seq analyses; Dr. Mario Suva and Dr. Birgit Knoechel for advice on lentiviral infections; Thanh Truong, Shawn Gillespie, and Kaylyn Williamson for help with reagents and ordering; and Dr. Mazhar Adli and Dr. Russell Ryan for conversation and advice.

I am also indebted to the Broad Institute Sequencing Platform and Broad Epigenomics Initiative: in particular, Tim Durham and Dr. Noam Shoresh for help with ChIP-seq and DamID-seq processing; and Dr. Charles Epstein, Robbyn Issner, Xiaolan Zhang, Mike Coyne, and Jeffrey Xing for help throughout the sequencing pipeline.

I would like to thank my Dissertation Advisory Committee, Prof. Ting Wu, Prof. Kami Ahmad, and Prof. Peter Park, for guiding and challenging me throughout the development of my Thesis. Thank you to my first rotation advisor, Prof. Martha Bulyk, and her former student Dr. Rachel Patton McCord for collaboration on the muscle enhancer project.

I acknowledge my fellowships that have supported me during all five years of study: the National Defense Science and Engineering Graduate Fellowship and the National Science Foundation Fellowship.

Furthermore, I would like to thank the current and former staff of the Biological and Biomedical Sciences office: Kate Hodgins, Maria Bollinger, Steve Obuchowski, and Daniel Gonzalez. I am also grateful for the staff at the Office of Career Services, particularly Dr. Laura Malisheski and Amy Sanford. Thank you to Dean John McNally and Dean David Cardozo for their strong support of graduate student life and insightful conversations and advice.

I also acknowledge the friends I have met through various graduate student organizations and committees, including the Healthcare Innovation and Commercialization Course, Harvard Graduate Consulting Club, DMS Alumni Committee, HILS Gala Welcome Dinner, Science in the News, and DMS Bulletin. I am also grateful for the great friends I have met during my five years in Boston. In particular, I would like to thank my dance partner of three years, Daniel Montana, for listening to the ups and downs of my graduate experience.

Finally, I owe much gratitude to my parents, Dr. Jerry Zhou and Yi Liu, and younger sister, Karen Zhou, for their continual love and support throughout my life.

Chapter 1:

Introduction –

Charting Histone Modifications and the

Functional Organization of Mammalian Genomes

Introduction –

Charting Histone Modifications and the

Functional Organization of Mammalian Genomes

This Chapter was published as:

"Charting Histone Modifications and the Functional Organization of Mammalian Genomes." (2011)

Nature Reviews Genetics 12(1):7-18

Vicky W. Zhou[1,2,3,4,*], *Alon Goren*[1,2,3,*] *and Bradley E. Bernstein*[1,2,3]

1. Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA, 02114.

2. Center for Systems Biology and Center for Cancer Research, Massachusetts General Hospital, Boston, Massachusetts, USA, 02114.

3. Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 02142.

4. Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA, 02115.

* These authors contributed equally to this work.

Correspondence to B.E.B.

e-mail: Bernstein.Bradley@mgh.harvard.edu

AUTHOR CONTRIBUTIONS

V.W.Z., A.G. and B.E.B. conceived and outlined the scope, content, and organization of this Review. V.W.Z. wrote the initial draft of all sections except “Replication time zones.” A.G. wrote the initial draft of the “Replication time zones” section. V.W.Z. drew all figures, which were then professionally improved by artists at *Nature Reviews Genetics*. V.W.Z., A.G., and B.E.B. edited the manuscript.

ABSTRACT

A succession of technological advances over the past decade have enabled researchers to chart maps of histone modifications and related chromatin structures with increasing accuracy, comprehensiveness and throughput. The resulting datasets highlight interplay between chromatin and genome function, dynamic variations in chromatin structure across cellular conditions, and emerging roles for large-scale domains and higher-ordered chromatin organization. Here we review a selection of recent studies that have probed histone modifications and successive layers of chromatin structure in mammalian genomes.

INTRODUCTION

The initial sequencing of the human genome a decade ago marked a shift away from a gene-centric paradigm and prompted many new lines of genome-scale investigation. An important emerging area relates to the packaging of DNA into chromatin and specifically how cell type-specific chromatin organization enables differential access and activity of regulatory elements and the manifestation of unique cellular phenotypes.

Eukaryotic chromatin structure can be viewed as a series of superimposed organizational layers (Felsenfeld and Groudine 2003; Schones and Zhao 2008) (Figure 1.1). At the root is the DNA sequence itself and its direct chemical modification by cytosine methylation (Law and Jacobsen). The DNA is folded into nucleosomes, the fundamental units of chromatin, that comprise approximately 147 bp of DNA wrapped around a histone octamer. The nucleosomal histones H2A, H2B, H3 and H4 can be chemically modified, and exchanged with variants.

Figure 1.1. Layers of chromatin organization in the mammalian cell nucleus.

Genomic DNA is methylated on cytosine bases in specific contexts, and is packaged into nucleosomes, which vary in histone composition and histone modifications; these features constitute the primary layer of chromatin structure. Here, different histone modifications are indicated by colored dots, and histone variants, such as H2A.Z, are shaded brown. DNA in chromatin may remain accessible to DNA-binding proteins such as transcription factors (TF) and RNA polymerase II (RNAPII), or may be further compacted. Chromatin can also organize into higher-order structures, such as lamina-associated domains and transcription factories. Each layer of organization reflects aspects of gene and genome regulation.

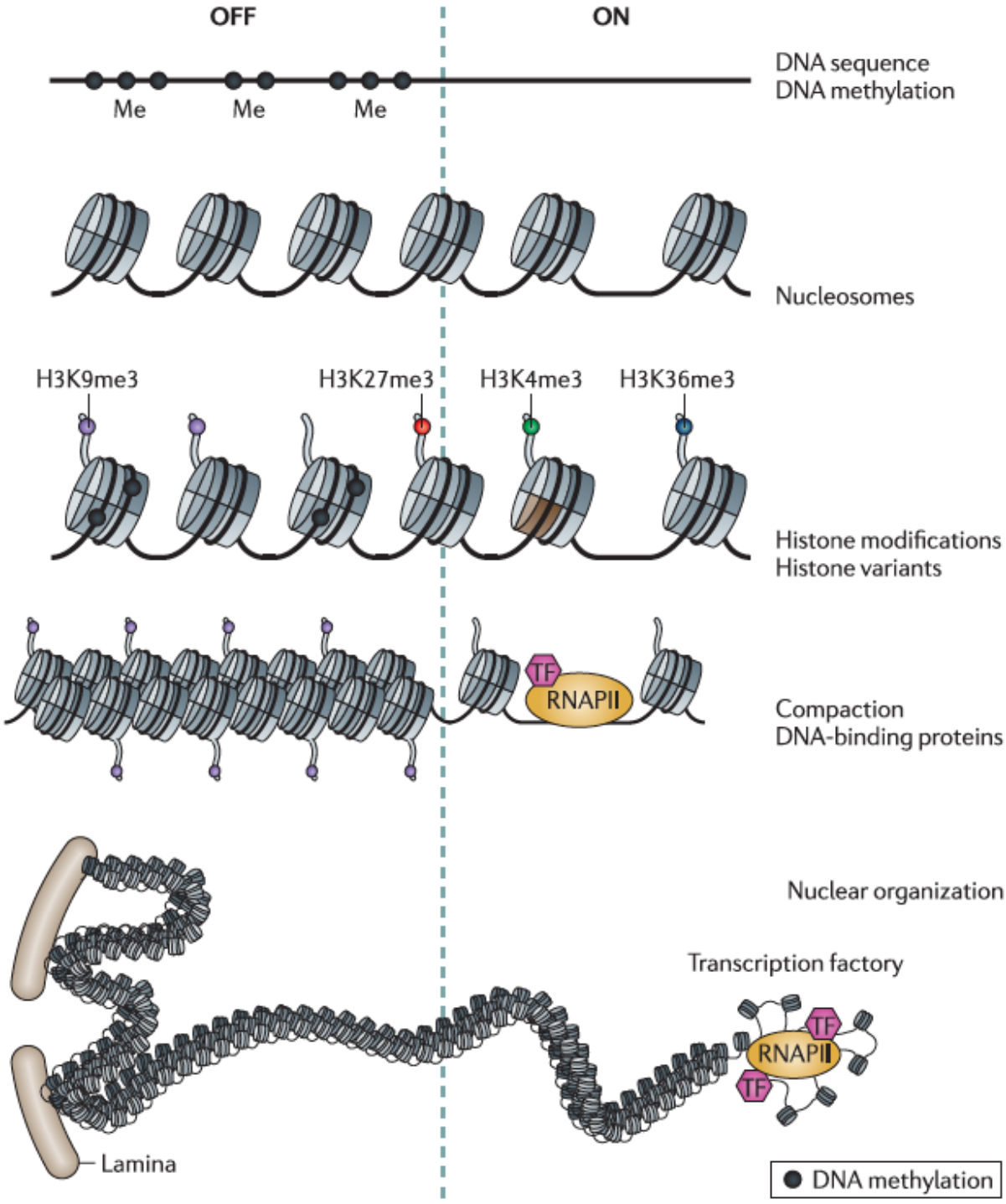

Figure 1.1 (Continued).

The nucleosome positions along with the variants and modifications make up the primary structure of chromatin. Finally, three-dimensional models of chromatin in the living cell are now being developed with increasing precision and propose additional sophisticated layers of regulation through higher-order organization and nuclear compartmentalization.

With increasing knowledge of chromatin structure and its attributes at different genomic loci and in various cell types comes the challenge to elucidate which elements and regulatory processes determine this structure. Specific chromatin configurations may be dictated by DNA sequence, DNA methylation patterns, transcription factors and other regulatory proteins, and transcriptional activity, and may be maintained through epigenetic controls rooted in the chromatin machinery (Margueron and Reinberg). Sequence features such as CpG islands, promoters and repetitive elements tend to assume characteristic modification patterns and chromatin states. These patterns result from complex mechanisms involving trans-acting factors that are subject to intense investigation but remain poorly understood (Margueron and Reinberg ; Bernstein, Meissner et al. 2007; Simon and Kingston 2009). These distinctive chromatin configurations facilitate targeting of transcription factors and regulatory machinery to active genomic elements within expansive mammalian genomes. As chromatin patterns at a particular locus are intimately related to underlying regulatory processes, they may vary markedly depending on cellular context. In particular, chromatin is heavily influenced by transcription factor networks and transcriptional processes which extensively harness chromatin modifiers and nucleosome remodelers (Li, Carey et al. 2007). In certain cases, environmental and stochastic events may invoke stable alterations in chromatin patterns, though our understanding of the output of such effects remains minimal (Jirtle and Skinner 2007).

Large-scale mapping of histone modifications and related structures has emerged as a

powerful means for characterizing the determinants as well as the functional consequences of chromatin structure. Here, we review recent studies that have applied technologies, such as ChIP-seq, to interrogate chromatin structure across the genome in diverse cell types, with an emphasis on mammalian models. We will briefly present the technological developments that have punctuated the shift from a gene-centric to genomewide view. Then we will discuss the current knowledge of the primary structure of chromatin, focusing on the global patterns, functions, and dynamics of histone modifications that overlay sequence features such as promoters, enhancers, and gene bodies. Finally, we will discuss notable recent studies that illuminate the link between histone modifications and higher-order chromatin domains.

FROM GENE-CENTRIC TO GENOMEWIDE

For the past several decades, chromatin biology has been guided by a succession of methodologies for probing features such as chromatin accessibility, DNA methylation, the locations, compositions and turnovers of nucleosomes, and the patterns of post-translational histone modifications. Technological advances in microarrays and next-generation sequencing have enabled many of these assays to be scaled genomewide. Notable examples include the DNaseI-seq (Boyle, Davis et al. 2008; Hesselberth, Chen et al. 2009), FAIRE-seq (Giresi, Kim et al. 2007), and Sono-seq (Auerbach, Euskirchen et al. 2009) assays for chromatin accessibility; whole genome and reduced-representation bisulfite sequencing (BS-seq) (Meissner, Mikkelsen et al. 2008; Lister, Pelizzola et al. 2009) and MeDIP-seq (Down, Rakyan et al. 2008) assays for DNA methylation; and the MNase-seq (Schones, Cui et al. 2008; Kaplan, Moore et al. 2009) and CATCH-IT (Deal, Henikoff et al.) assays for elucidating nucleosome positions and turnovers,

respectively. These technologies and their integration have been extensively reviewed elsewhere (Hawkins, Hon et al. ; Park 2009). In this section, we will focus on histone modifications and, in particular, on how genomewide ChIP-seq mapping studies have enhanced our understanding of the chromatin landscape.

Mapping histone modifications genomewide.

While chromatin immunoprecipitation (ChIP) has been used since 1984 (Gilmour and Lis 1984; Solomon and Varshavsky 1985) to probe chromatin structure at individual loci, its combination with microarrays, and more recently next-generation sequencing, has advanced far more precise and comprehensive views of the modification landscapes, highlighting roles for chromatin structures across diverse genomic features and elements not appreciated in targeted studies. The basis of ChIP is the immunoprecipitation step in which an antibody is used to enrich chromatin that carries a histone modification (or other epitope) of interest. In ChIP-seq, next-generation technology is used to deep sequence the immunoprecipitated DNA molecules and thereby produce digital maps of ChIP enrichment. An example is the comprehensive work by the Zhao group to profile 39 different histone methylation and acetylation marks genomewide in human CD4+ T cells (Barski, Cuddapah et al. 2007; Wang, Zang et al. 2008). These maps and similar datasets (Guenther, Levine et al. 2007; Heintzman, Stuart et al. 2007; Mikkelsen, Ku et al. 2007) have associated particular modifications with gene activation or repression and with various genomic features, including promoters, transcribed regions, enhancers, and insulators (Figure 1.2).

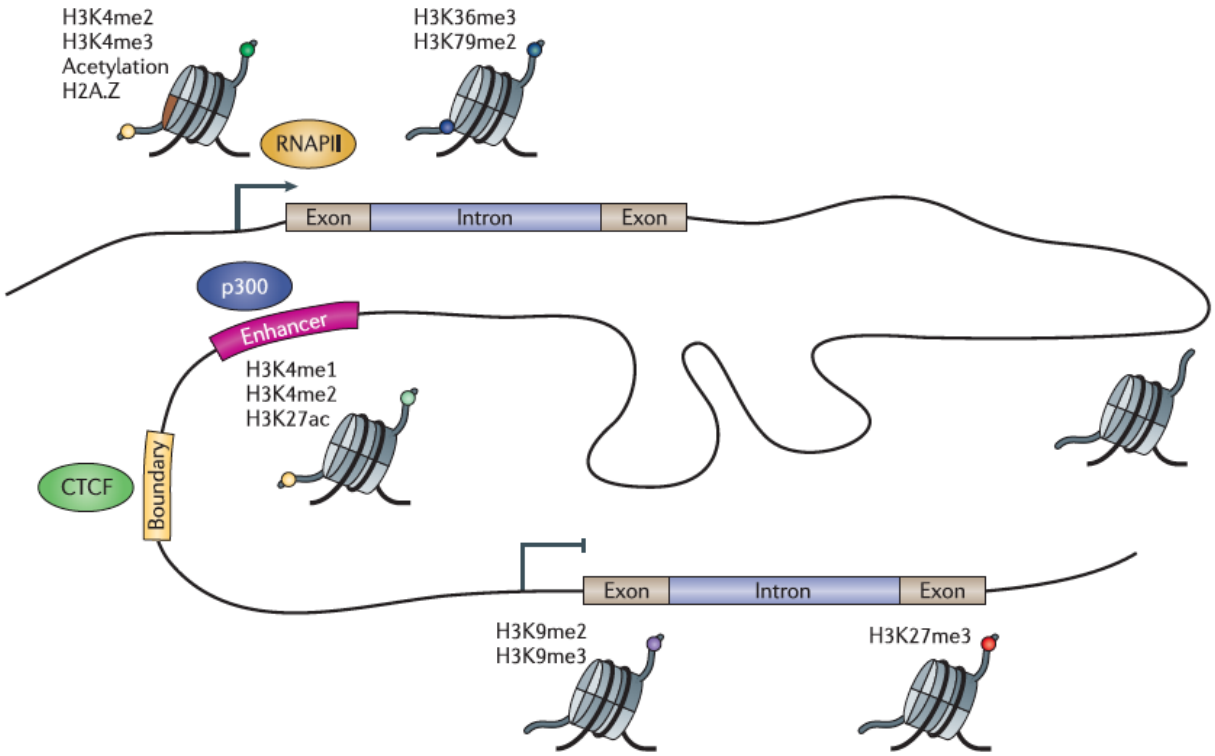

Figure 1.2. Histone modifications demarcate functional elements in mammalian genomes. Promoters, gene bodies, an enhancer, and a boundary element are indicated on the representative genomic interval. Active promoters are commonly marked by H3K4me2/3, acetylation, and H2A.Z. Transcribed regions are enriched for H3K36me3 and H3K79me2. Repressed genes may be located within large domains of H3K9me2/3 or H3K27me3. Enhancers are relatively enriched for H3K4me1/2, H3K27ac, and p300. The CTCF protein binds many sites that may function as boundary elements, insulators or structural scaffolds. These features organize the DNA and distinguish functional elements within the large expanse of genome.

Box 1.1. ChIP-seq: Current limitations and future progress.

Enabled by technological advances and plummeting costs of DNA sequencing, genomewide maps for histone modifications and related chromatin structures are being generated at ever increasing rates. Given this expanding reliance on ChIP-seq technology and data, there is a need for uniform implementation of data standards. The ENCODE Project (Birney, Stamatoyannopoulos et al. 2007) and the NIH Roadmap for Epigenomics (Bernstein, Stamatoyannopoulos et al. 2010) have established standards for experimental procedures, documentation and quality controls intended to ensure the quality and facilitate the portability, interpretation and integration of functional genomic data.

Questions also remain at the level of biological interpretation of ChIP-seq data. Inherent to ChIP technology is that it reports on the relative enrichment of a modification across a population of cells. Accordingly, it cannot discern the absolute level of these modifications, *i.e.* what fraction of histone tails at a given locus are modified, and may be confounded by cellular heterogeneity. The magnitude of enrichment signal is also an important consideration. A select few modifications typically show enrichments of 10- to 100-fold and thereby offer particularly reliable metrics. Signals for many other epitopes tend to be subtler, but may be equally biologically important. In such cases, it can be difficult to distinguish whether perceived differences reflect technical issues such as inefficient immunoprecipitation, or true biological phenomena. Statistically significant trends can often be discerned through composite analysis of hundreds of genes or elements, but biological conclusions should be made with care when overall magnitude differences are incremental. Although these limitations are starting to be addressed by improved ChIP-seq procedures that increase sensitivity and reliability, there is an urgent need for orthogonal approaches.

These and subsequent studies highlight the value of the comprehensive and less biased sequencing approaches for testing the generality of insights gleaned through gene-specific studies as well as for identifying altogether new associations and biological phenomena (Box 1.1).

Integrating ChIP-seq maps.

The expanding body of chromatin data in the public domain has fostered many computational efforts aimed at integrating different data types, identifying novel relationships among histone modifications and related chromatin structures, and developing new hypotheses regarding their regulatory functions. Integrations of histone modification maps with chromatin accessibility, nucleosome positions, transcription factor binding, RNA expression and sequence-based genome annotations are providing increasingly unified views of chromatin structure and function (Hawkins, Hon et al. ; Birney, Stamatoyannopoulos et al. 2007; Kaplan, Moore et al. 2009).

Two recent studies present innovative approaches for integrating genomewide chromatin maps (Hon, Wang et al. 2009; Ernst and Kellis), both of which were demonstrated on a compendium of ChIP-seq data for human CD4+ T cells (Barski, Cuddapah et al. 2007; Wang, Zang et al. 2008). Hon *et al.* applied a pattern-finding algorithm called ChromaSig to identify combinations of histone modifications at predetermined classes of regulatory loci, including promoters and enhancers. After validating that their approach identified known associations between modifications and expression levels, they applied it to regions outside of these elements and subsequently identified distinct chromatin signatures associated with exons and large-scale

repressed regions. Ernst *et al.* used a multivariate Hidden Markov Model to discover biologically meaningful combinations *a priori*. They discovered 51 distinct chromatin states that could be sub-divided according to current genome annotations, including several promoter-associated, enhancer-associated and repressed states. This unbiased approach revealed the high information content provided by combinatorial modification patterns. It also confirmed striking functional distinctions between methylation marks affecting different residues or assuming different degrees of chemical modification (mono-, di- or tri-methylation). In contrast, the functional correlates of histone acetylation marks appeared to be less dependent on the specific residues involved, but rather on the overall degree of acetylation, consistent with previous studies in yeast (Durrin, Mann et al. 1991; Dion, Kaplan et al. 2007).

Although their findings are largely consistent with prior knowledge of histone modification functions, these studies are significant for their forward-looking approaches to developing algorithms that integrate increasingly vast bodies of functional genomic data into coherent biological views. An important future direction will be an equally systematic characterization of chromatin-associated proteins, including the regulators that modify and otherwise interact with histones. Such data could facilitate perturbation of specific chromatin structures to thereby yield insight into their functions. This goal will be challenged by technical difficulties. However, a recent study in *Drosophila* that localized dozens of chromatin proteins, and thereby partitioned the genome based on their combinatorial binding patterns, provides a potential path forward (Filion, van Bemmel et al. 2010).

HISTONE MODIFICATIONS ACROSS GENOMIC SEQUENCE ELEMENTS

In this section, we will review the types and patterns of histone modifications that have been linked to major functional genomic elements, discuss their dynamics through cell differentiation and development, and touch upon functional studies that are beginning to give a mechanistic grounding to these observed patterns.

High and low CpG content promoters (HCPs and LCPs).

Although mammalian promoter regions vary considerably in their positional relationships to genes, the DNA sequence proximal to the transcriptional start site (TSS) of a gene (e.g., the region +/- 2 kb) is frequently considered as a proxy. The patterns of histone modification across such regions offer insight into the regulatory state of promoters and genes, and have revealed important paradigms of gene regulation.

Mammalian promoters can be classified according to their sequence content, and this has proven to be useful for understanding their regulation (Figure 1.3). A majority of promoters coincide with regions of high GC content and CpG ratios, or 'CpG islands'. These have been termed 'high CpG content promoters' or 'HCPs', in contrast to 'low CpG content promoters' or 'LCPs'. HCPs and LCPs have different histone modification patterns and distinct modes of regulation (Mikkelsen, Ku et al. 2007; Weber, Hellmann et al. 2007). The distinction between HCPs and LCPs is somewhat arbitrary and does not effectively address a number of intermediate CpG content promoters.

Figure 1.3. Chromatin patterns and regulation by promoter class.

Promoters may be classified according to their CpG content. High CpG and low CpG content promoters, HCPs and LCPs, respectively, are subject to distinct chromatin patterns and regulation. (a) HCPs have characteristics of accessible or 'active' chromatin by default. Active HCPs (*e.g.* housekeeping gene promoters) are enriched for H3K4me3 and subject to RNAPII initiation. They may be subject to additional regulation at the transition to elongation. (b) Poised HCPs (*e.g.* developmental regulator gene promoters in ES cells) are marked by the bivalent combination of H3K4me3 and H3K27me3. They may be subject to RNAPII initiation, but tend not to elongate or make productive mRNA. (c) Inactive HCPs carry 'repressive' chromatin modifications such as H3K27me3, and are relatively inaccessible to RNAPII. Unlike HCP chromatin, LCP chromatin appears to be selectively activated (e.g., by specific transcription factors or 'TFs'). (d) Active LCPs are enriched for H3K4me3 and transcribed. (e) Poised LCPs may be marked by H3K4me2 without H3K4me3. (f) Inactive LCPs typically lack chromatin marks but may be DNA methylated.

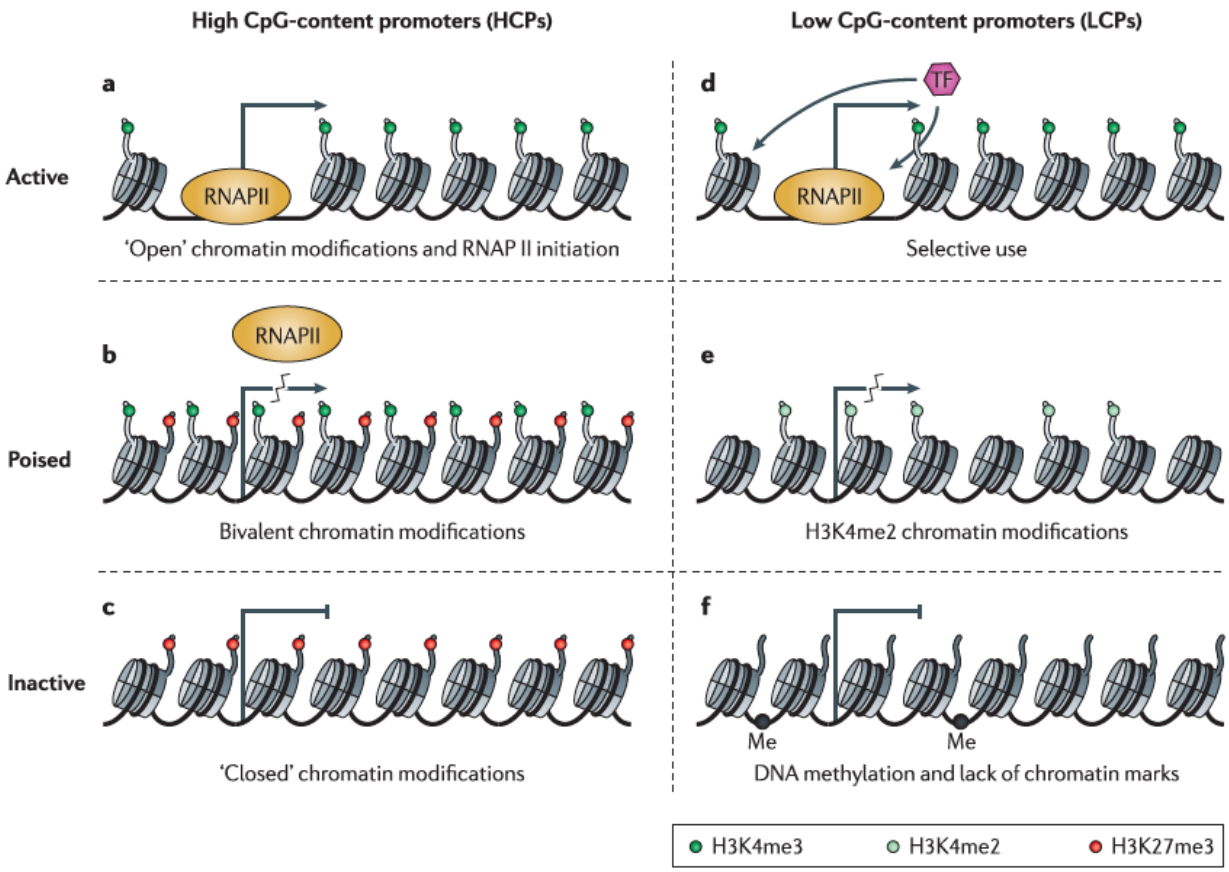

Figure 1.3 (Continued).

Incorporation of additional sequence features such as DNA motifs or DNA methylation patterns may result in a more precise and biologically-meaningful classification (Bernstein, Meissner et al. 2007; Straussman, Nejman et al. 2009). Nonetheless, the two classes provide a useful framework for understanding and distinguishing the functions and regulation of mammalian promoters.

Initial ChIP-chip studies in mammalian cells revealed punctate peaks of H3K4me3 in association with TSSs of many transcribed genes (Bernstein, Kamal et al. 2005; Kim, Barrera et al. 2005) (Figure 1.3). Subsequent studies of embryonic stem (ES) cell chromatin revealed surprisingly broad targeting of H3K4me3 to virtually all HCPs, regardless of expression state (Guenther, Levine et al. 2007; Mikkelsen, Ku et al. 2007). Sites of H3K4me3 were shown to be accompanied by other features of accessible chromatin, including histone acetylation, occupancy by the H3.3 histone variant and hyper-sensitivity to DNase I digestion (Goldberg, Banaszynski et al. ; Wang, Zang et al. 2008; Hon, Wang et al. 2009; Ernst and Kellis). Differentiated cells were also found to exhibit relatively broad targeting of H3K4me3 to promoters, though with specific and biologically-meaningful exceptions (Mikkelsen, Ku et al. 2007) (see *Poised and repressed chromatin states*, below).

These accessible, H3K4me3-marked regions are also hypo-methylated at the DNA level, as expected from their high CpG contents (Weber, Hellmann et al. 2007; Meissner, Mikkelsen et al. 2008). This is consistent with a general exclusivity between such activating and 'open chromatin' structures and DNA methylation. Indeed, several studies have provided evidence for direct antagonism between these epigenomic features. For instance, methylation of H3K4 was shown to preclude a physical interaction between the histone tail and DNMT3L (Ooi, Qiu et al. 2007). Another study, in the plant *Arabidopsis thaliana*, reported a direct role for H2A.Z – a

histone variant enriched in actively exchanging chromatin – in protecting gene promoters from DNA methylation. In addition to a global exclusivity between sites of H2A.Z deposition and DNA methylation, this study also demonstrated that deficiency of H2A.Z deposition led to generalized DNA hyper-methylation (Zilberman, Coleman-Derr et al. 2008).

What mechanisms could underlie the correspondence between these 'open chromatin' features, H3K4me3 and the GC-rich promoters? ChIP-chip studies in ES cells showed that many H3K4me3-marked promoters are also enriched for RNA polymerase II (RNAPII) and subject to transcriptional initiation (Guenther, Levine et al. 2007). This was a surprising finding given that a substantial fraction of the HCPs do not make productive RNA transcripts or even undergo elongation (see *Poised and repressed chromatin states*, below). It suggests that transcriptional initiation and H3K4me3 are tightly linked and, moreover, that initiating RNAPII substantially contributes to the accessible chromatin configuration potentially through interactions with chromatin modifiers as seen in yeast (Li, Carey et al. 2007; Shilatifard 2008). The concordance between H3K4me3 and HCPs may be more directly explained by physical recognition of un-methylated CpG dinucleotides by CXXC domains in H3K4 methyltransferase complexes (Lee and Skalnik 2005). It was recently shown that introducing artificial, promoterless, CpG clusters into mouse ES cells was sufficient to recruit the Set1 complex and establish H3K4me3 (Thomson, Skene et al.). A parallel study demonstrating targeting of an H3K36 demethylase complex by its CXXC domain suggests that such interactions may be general (Blackledge, Zhou et al. 2010). Together, these converging lines of experimental evidence suggest that transcriptional initiation and alternate pathways mutually reinforce a chromatin configuration that distinguishes this essential promoter class.

Regardless of the relative contributions of these models, the data suggest that HCPs tend

to adopt an accessible chromatin state by default, and are generally subject to a degree of initiation. Thus, effective regulation of HCP genes likely requires additional controls. Indeed, recent studies in macrophages and ES cells have documented roles for specific transcription factors in regulating steps downstream of initiation (Rahl, Lin et al. ; Hargreaves, Horng et al. 2009; Ramirez-Carrozzi, Braas et al. 2009). The Smale and Medzhitov groups characterized a class of HCPs with constitutively active chromatin in macrophages that are basally transcribed by RNAPII, generating non-functional RNAs. After the macrophages were induced by lipopolysaccharide (LPS), the transcription factor NF-KB initiates a cascade that causes RNAPII to adopt a more processive form (*i.e.* from serine 5 to serine 2 phosphorylation on its C-terminal domain) and results in rapid production of functional transcripts. In ES cells, genomewide mapping studies revealed a key role for the transcription factor c-Myc in enhancing the 'release' of RNAPII at HCPs, and hence promoting the generation of mature transcripts. Together, these studies emphasize the importance and complexity of downstream steps in controlling the expression of genes associated with this major promoter class.

In marked contrast to HCPs, LCPs appear to be inactive by default (Figure 1.3). Indeed, most annotated LCPs lack H3K4me3 (or H3K4me2) in ES cells as well as in various differentiated cell types (Mikkelsen, Ku et al. 2007; Weber, Hellmann et al. 2007).The minority of LCPs that are marked by H3K4me3 appear to be fully expressed with RNA levels substantially higher than their unmarked counterparts.

Further biological insight into LCP regulation emerged from an analysis of chromatin structure changes during hematopoietic differentiation (Orford, Kharchenko et al. 2008). Orford *et al*. defined a subset of promoters that carry H3K4me2 but not H3K4me3 in hematopoietic progenitors. They found that this set corresponded to LCPs associated with hematopoietic cell

type-specific genes that are generally inactive in the progenitors but become induced during differentiation. They specifically observed a switch from H3K4me2 to H3K4me3 upon induction of such LCPs during differentiation. These studies suggest that LCPs are subject to greater regulation at the level of initiation, and may be poised in certain contexts by lower degrees of histone methylation. Notably, genes subject to this form of regulation tend to encode terminal cell type-specific factors (e.g., structural proteins) as opposed to the master regulators that drive cell fate (e.g., developmental transcription factors). The latter have HCPs and are subject to more complex regulation by Polycomb complexes (see below).

Poised and repressed chromatin states.

Repressed promoters also exhibit unique patterns of chromatin modifications that appear to reflect distinct modes of transcriptional silencing. These include H3K27me3, the proto-typical mark of Polycomb repressors; H3K9me3, which correlates with constitutive heterochromatin; and DNA methylation (Figure 1.4a).

Polycomb proteins are transcriptional repressors essential for maintaining tissue-specific gene expression programs in multi-cellular organisms (Simon and Kingston 2009). In mammals, a large proportion of HCPs are targeted by the main Polycomb repressive complexes, PRC1 and PRC2. Roughly 20% of HCPs are bound by PRC2 and marked by the associated modification, H3K27me3 (Boyer, Plath et al. 2006; Lee, Jenner et al. 2006; Mikkelsen, Ku et al. 2007; Pan, Tian et al. 2007; Zhao, Han et al. 2007; Ku, Koche et al. 2008). These promoters have been termed ‘bivalent’ as they also carry H3K4me3 and thus have characteristics of both activating and repressive chromatin (Azuara, Perry et al. 2006; Bernstein, Mikkelsen et al. 2006).

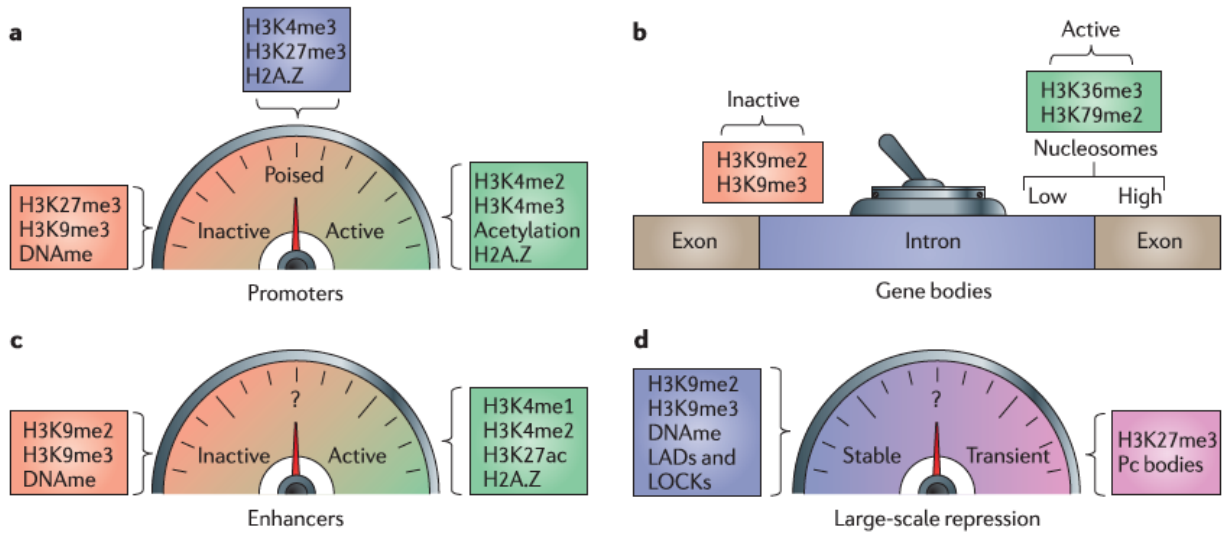

Figure 1.4. "Dashboard" of histone modifications for fine-tuning genomic elements.

In addition to enabling annotation, histone modifications may serve as 'dials' or 'switches' for cell type specificity. (a) At promoters, they can contribute to fine-tuning of expression levels, from on to poised to off, and perhaps even intermediate levels. (b) At gene bodies, they discriminate active and inactive conformations. In addition, exons in active genes have higher nucleosome occupancy, and thus more H3K36me3 and H3K79me2-modified histones, than introns. (c) At distal sites, they correlate with levels of enhancer activity. (d) At a global scale, they may confer repression of varying stabilities.

Bivalent, PRC2-target promoters have attracted considerable interest as a large proportion corresponds to developmental genes encoding transcription factors and other regulators of cellular state. These genes are largely inactive in pluripotent cells, but subsequently may be rapidly induced or stably inactivated, depending on the developmental course. It has been proposed that the signature chromatin configuration is instrumental for poising bivalent promoters for these alternate fates. Indeed, global studies of neural and hematopoietic progenitors indicate that bivalent chromatin tends to resolve at successive developmental stages in a pattern that closely matches the expression state and future potential of the corresponding genes (Mikkelsen, Ku et al. 2007). For example, Mohn *et al.* followed H3K27me3 patterns in gene promoters during transition of ES cells to neural progenitors and subsequently to terminal neurons, finding progression of HCP modifications in accordance with expression state and gene potential (Mohn, Weber et al. 2008). Similar patterns are also evident along the axis of hematopoietic differentiation, as indicated by analysis of *in vivo* lineages from both human and mouse (Adli, Zhu et al. ; Cui, Zang et al. 2009).

Although bivalent promoters in ES cells have very low expression levels and were initially found to be free from RNAPII (Mohn, Weber et al. 2008), subsequent studies have suggested that at least a subset do have detectable RNAPII enrichment (Guenther, Levine et al. 2007; Stock, Giadrossi et al. 2007). This raises the possibility that initiating RNAPII contributes to the establishment of H3K4me3, or potentially even H3K27me3 at these loci. Still, the data should be interpreted with some caution. RNAPII enrichment was only detected under certain experimental conditions (Stock, Giadrossi et al. 2007) and, moreover, evidence for RNA transcription at these loci remains scarce (Seila, Calabrese et al. 2008). Other technical issues of possible relevance include an inherent promoter bias in some ChIP data, and heterogeneity due

to partial differentiation.

How is PRC2 targeted to HCPs? The GC-rich sequence is likely to play an important role here given the extremely high correspondence between CpG islands and PRC2 binding. PRC2 targets in ES cells can be predicted with remarkable accuracy by simply identifying CpG islands that lack motifs for activating transcription factors (Ku, Koche et al. 2008). A causal role for such sequences is supported by the finding that introduction of exogenous GC-rich sequence elements into ES cells is sufficient to mediate PRC2 recruitment (Mendenhall, Koche et al. 2010). Still, the underlying mechanisms are not yet understood. Although sequence-specific DNA binding proteins guide PRC2 to target elements in *Drosophila*, analogous factors have yet to be demonstrated in mammals. Rather, mammalian PRC2 contains the atypical DNA binding proteins Aebp2 (Kim, Kang et al. 2009) and Jarid2 (Li, Margueron et al. ; Pasini, Cloos et al. ; Peng, Valouev et al. 2009; Shen, Kim et al. 2009). Jarid2 was recently shown to be essential for PRC2 function and the establishment of proper K27me3 patterns (Li, Margueron et al. ; Pasini, Cloos et al. ; Peng, Valouev et al. 2009; Shen, Kim et al. 2009). ChIP-seq analysis confirmed that Jarid2 co-localizes with PRC2 and K27me3 at GC-rich sequence elements. However, *in vitro* biochemical studies suggest that Jarid2 is a promiscuous DNA binding protein without particular specificity for GC-rich sequences (Kim, Kraus et al. 2003). Hence, this factor does not appear to fully explain PRC2 targeting. Non-coding RNAs have also emerged as intriguing candidates for recruitment. PRC2 has affinity for a variety of RNA classes, such as short GC-rich RNAs that might play a role in targeting the complex to weakly initiating HCPs (Kanhere, Viiri et al.). The complex can also interact with long intergenic non-coding RNAs (lincRNAs), including Xist and HOTAIR, both of which appear to play important roles in the localization and stabilization of Polycomb complexes in differentiating cells (Tsai, Manor et al. ; Zhao, Sun et al.

2008). PRC2 association is further stabilized by the innate affinity of the complex for K27-methylated H3 tails (Hansen, Bracken et al. 2008; Margueron, Justin et al. 2009). Thus, in contrast to *Drosophila*, PRC2 localization in mammals appears to be directed to GC-rich elements by a complex interplay between low-specificity DNA binding proteins, RNA targeting factors and chromatin-based stabilization.

The challenge of understanding Polycomb localization is further complicated by PRC1, an additional repressive complex that ubiquitinylates histone H2A and may also mediate structural compaction of chromatin (Simon and Kingston 2009). In ES cells, PRC1 associates with a specific subset of PRC2 targets that includes key developmental regulators and other genes subject to epigenetic repression during development (Ku, Koche et al. 2008). These PRC1 targets tend to have larger CpG islands or extended GC-rich regions relative to PRC2-specific loci. In addition, recent studies have identified specific DNA elements with YY1 motifs capable of initiating PRC1-dependent silencing during development (Woo, Kharchenko et al. ; Sing, Pannell et al. 2009). A unifying theory for the complexes is an important goal as both PRC1 and PRC2 are almost certainly required for stable epigenetic gene repression (Simon and Kingston 2009).

The landscape of Polycomb repression changes markedly through differentiation. In addition to the progressive resolution of bivalent chromatin at specific promoters described above, a smaller subset of promoters is subject to *de novo* gain of H3K27me3 during development (Mohn, Weber et al. 2008). The affected genes include certain pluripotency regulators repressed during ES cell differentiation (Pan, Tian et al. 2007). At many loci, differentiation is also accompanied by dramatic spreading to yield contiguous if more diffuse domains of H3K27me3 (Hawkins, Hon et al.).

Relatively less is known about the role of DNA methylation in HCP regulation during development. Hyper-methylation of individual CpG islands as well as extended genomic loci has been widely described in human cancer (Coolen, Stirzaker et al. ; Jones and Baylin 2007). Yet genome-scale studies suggest that most CpG islands remain largely un-methylated during normal development (Meissner, Mikkelsen et al. 2008; Mohn, Weber et al. 2008). Still, a closer look at the DNA methylation pattern of HCPs shows that even though the CpG islands are un-methylated, their 'shores' – sequences up to 2kb distant from the CpG islands – frequently become methylated in tissue-specific patterns (Irizarry, Ladd-Acosta et al. 2009). CpG island shores may also be conserved between human and mouse, and when methylated correlate to gene silencing in a tissue specific manner. Although the functionality of the shores remains controversial, global reduction of DNA methylation by a small molecule (5-AZA-c) or by knockout of DNA methyltransferases shows concurrent activation of these genes. More broadly, genome-scale and genomewide analysis of DNA methylation patterns have provided insight into ES cell regulation (Lister, Pelizzola et al. 2009), hematopoietic differentiation (Ji, Ehrlich et al.), and epigenetic roadblocks to cellular reprogramming (Mikkelsen, Hanna et al. 2008).

Up to 80% of LCPs are DNA methylated in ES cells (Fouse, Shen et al. 2008; Meissner, Mikkelsen et al. 2008). The functional consequence of the DNA methylation remains unclear – the relative paucity of CpG dinculeotides in these regions suggests that the impact may be slight. Interestingly, however, inactive LCPs are frequently located within extended regions of H3K27me3 or H3K9me3 that may reflect large-scale sequestration of inactive genomic regions, and thereby hold potential for contextual repression of chromosomal regions (see *H3K9me3 and lamina-associated domains*, below).

Gene bodies.

Mammalian genes are characterized by large numbers of exons in an expanse of introns. In many cases, alternative splicing provides an additional layer of complexity and regulation (Nilsen and Graveley). Recent studies suggest that chromatin patterns can distinguish primary transcripts and exons, and may even play a role in determining splicing patterns. Major marks seen in transcribed regions include H3K36me3 (Barski, Cuddapah et al. 2007; Mikkelsen, Ku et al. 2007) and H3K79me2 (Li, Carey et al. 2007) (Figure 1.4b). Comparative analyses of H3K36me3 with expression and splicing data reveal several emerging trends. First, H3K36me3 levels correlate with levels of gene expression (Barski, Cuddapah et al. 2007; Mikkelsen, Ku et al. 2007). This likely reflects interactions between elongating RNAPII and the corresponding methyltransferases (Li, Carey et al. 2007).

Recent studies have noted that expressed exons, as opposed to introns, have particularly strong enrichment for H3K36me3 (Andersson, Enroth et al. 2009; Kolasinska-Zwierz, Down et al. 2009; Schwartz, Meshorer et al. 2009). They may also exhibit modest enrichment for H2BK5me1, H4K20me1, and H3K79me1 (Ernst and Kellis). Subsequent studies have indicated that the observed enrichments for histone marks likely reflect preferential occupancy and positioning of nucleosomes over exons (Schwartz, Meshorer et al. 2009; Tilgner, Nikolaou et al. 2009) (Figure 1.4b). Specifically, computational analyses in the latter studies suggest that this higher abundance of nucleosomes might account for the observed exonic H3K36me3 enrichment. The authors of these studies speculated that positioned nucleosomes at exons might enhance splicing by acting as ‘speed bumps’ to slow RNAPII. According to this model, the

splicing machinery is recruited during transcription, and an increased RNAPII occupancy time might translate into improved recognition of splicing signals (Kornblihtt, Schor et al. 2009).

A recent study by the Misteli group more directly linked histone modifications at gene bodies with the splicing machinery (Luco, Pan et al.). These authors studied the alternatively spliced gene, FGFR2. They found that histone modifications across the gene vary between cell types. Specifically, they observed distinct patterns of H3K36me3, H3K4me3, H3K4me1 and H3K27me3 over FGFR2 in epithelial cells and mesenchymal cells, which produce different splice forms. Remarkably, by modulating the levels of H3K36me3 and H3K4me3, the authors were able to influence the splicing patterns. They suggest a model in which histone marks are read by the splicing machinery via the histone-tail binding protein MRG15 and the splicing regulator PTB. Interestingly, if these histone patterns are general signatures of alternatively spliced exons, a comparison of genomewide maps of these marks in different cell types may reveal global maps of alternative splicing events. Regardless, the robust enrichment of modified nucleosomes at exons suggests that a link between histone modifications and splicing may be a general phenomenon.

Enhancers.

Enhancers are DNA elements that recruit transcription factors, RNAPII, and chromatin regulators to positively influence transcription at distal promoters (Visel, Rubin et al. 2009). Histone modification profiles have proven to be particularly useful for identifying enhancer elements in unbiased fashion. In addition to specific histone modifications, these elements are preferentially occupied by sequence-specific DNA binding proteins (Birney,

Stamatoyannopoulos et al. 2007) and coactivators such as p300 (Visel, Blow et al. 2009) (Figure 1.2). By observing the histone modifications at distal p300-binding sites, Heintzman *et al.* identified relative H3K4me1 enrichment and relative H3K4me3 depletion as a chromatin signature of enhancers in human cells (Heintzman, Stuart et al. 2007). The group used this signature to predict over 55,000 candidate enhancers in five human cell types, including K562 and HeLa (Heintzman, Hon et al. 2009). Remarkably, the chromatin patterns at enhancers were much more variable and cell type-specific than chromatin patterns at promoters or insulators. This study suggested a critical role for enhancers in controlling the level and timing of expression in a cell type-specific manner and highlights the power of histone modification profiling for capturing diverse functional elements.

Despite the fruitful application of a histone modification signature to predict enhancers, the mechanism by which H3K4me1 is established at these sites remains unknown. Integrative analyses suggest that enhancers also share enrichment for H3K27 acetylation, H2BK5me1, H3K4me2, H3K9me1, H3K27me1, and H3K36me1, suggesting redundancy in the histone marks (Hon, Wang et al. 2009). This may be an indication of generalized accessibility or chromatin dynamics at these sites. It may also reflect physical proximity of the enhancer elements to activating chromatin machinery at their target promoters, through looping interactions (Visel, Rubin et al. 2009). The chromatin patterns at enhancers may also be actively fine-tuned as different patterns of acetylation and H2A.Z deposition correlate with differences in downstream gene expression levels (Ernst and Kellis) (Figure 1.4c).

Support for a more direct interaction between enhancers and the transcriptional machinery emerged from a recent genomewide study that mapped p300 and H3K4me1 in mouse cortical neurons. Kim *et al.* found that RNAPII interacts with many active enhancers identified

by the chromatin patterns in these cells and transcribes bi-directional short (<2kb) non-coding RNAs, termed eRNAs (Kim, Hemberg et al.). The expression levels of eRNAs correlate with the proximal gene activity, and eRNA synthesis appeared to require interaction with the relevant promoter. The exact function of these eRNAs is not understood, but similar findings also emerged from a study of enhancer elements in macrophages (De Santa, Barozzi et al. 2010). Transcription of eRNAs could be needed to maintain open chromatin at the enhancer region but, alternatively, may be a by-product of the chromatin configuration or looping.

Insulators and boundary elements.

Insulators are DNA elements that block enhancer activities (Phillips and Corces 2009) (Figure 1.2). They are likely related to boundary elements defined by their capacity to prevent heterochromatin spreading. In mammals, the CTCF protein has been both implicated in these two processes, and in inter- and intra-chromosomal organization. CTCF has been profiled genomewide in several human cell types, revealing tens of thousands of binding sites in primary human fibroblasts, CD4+ T cells, and HeLa cells (Barski, Cuddapah et al. 2007; Kim, Abdullaev et al. 2007; Heintzman, Hon et al. 2009). These studies come to the consensus that the majority of CTCF binding sites share a common motif and are relatively invariant across different cell types. The CTCF binding sites also show modest enrichment for the histone variant H2A.Z, but surprisingly, vary widely in terms of other accompanying histone modifications (Hon, Wang et al. 2009; Ernst and Kellis). Recent models suggest that CTCF, most likely in association with cohesion (Wendt, Yoshida et al. 2008), stabilizes long-range DNA interactions and chromatin loops. In this way, the factor is thought to be instrumental in establishing a defined three-

dimensional genome structure and partitioning distinct chromatin domains (Phillips and Corces 2009).

HIGHER-ORDER CHROMATIN ORGANIZATION

As cells differentiate from a totipotent to a specialized committed state, a high percentage of their genome must be stably repressed. In this regard, chromatin regulators and histone modifications appear to work in conjunction with other mechanisms to silence broad genomic regions. There are several known modes of large-scale repression that correlate with megabase (Mb) domains of H3K9me3 and H3K27me3 and likely reflect specialized higher-order chromatin structures within the nucleus (Figure 1.4d, Figure 1.5).

H3K9me3 and lamina-associated domains.

The nuclear lamina is thought to bind and silence large regions of heterochromatin. Two studies probing for distinct genomic features identified similar sets of domains enriched for H3K9 methylation and lamina contact (Guelen, Pagie et al. 2008; Wen, Wu et al. 2009). Guelen *et al.* globally mapped the interaction between the genome and nuclear lamina in human fibroblasts using DamID. These authors observed two discrete chromatin environments: lamina-associated domains (LADs), and regions outside LADs. Both regions were on the order of 0.1 – 10 Mb in size. LADs were found to have low gene density, low transcriptional activity and a paucity of active chromatin modifications.

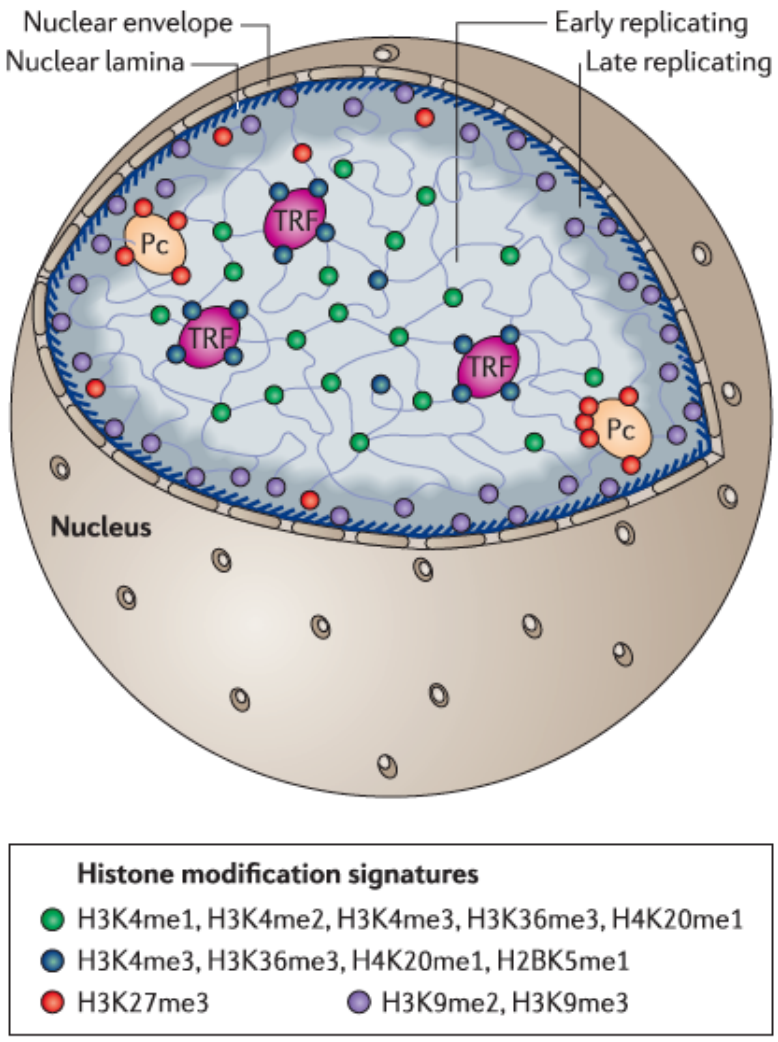

Figure 1.5. Histone modification signatures associated with features in the mammalian cell nucleus.

Signature histone modifications correlate with various nuclear features, although they may not have a one-to-one correspondence. Chromatin with generally active modifications (green dots) often replicates early, while chromatin with generally repressive modifications (purple dots) replicates late. Regions enriched for active modifications (blue dots) may converge into transcription factories (TF). Blocks of H3K27me3 (red dots) may form Polycomb bodies (Pc), while diffuse domains marked by H3K9me3 (orange dots) appear to contact the nuclear lamina.

Although the nuclear lamina had previously been associated with inactivity, these studies defined for the first time their locations, extents and correlated chromatin patterns. Remarkably, tethering experiments show that interaction with the nuclear lamina is not only correlative but is also causal in reducing gene expression (Finlan, Sproul et al. 2008; Kumaran and Spector 2008; Reddy, Zullo et al. 2008).

Wen *et al.* identified a similar set of genomic domains by analyzing genomewide maps of H3K9me2 in differentiated and undifferentiated cells (Wen, Wu et al. 2009). They found large and diffuse regions of K9 methylation that cover up to 4.9 Mb and collectively represent up to 46% of the genome, which they termed LOCKs. These investigators also showed that LOCKs are conserved between human and mouse, and that the H3K9me2 was dependent on the G9a H3K9 methyltransferase. Furthermore, a close relationship between LOCKs and LADs was indicated by a striking overlap of 82% between placental LOCKs and LADs found in fibroblasts. Thus, genomic regions diffusely marked by H3K9 methylation seem to be in contact with the nuclear lamina; these findings have prompted a model in which chromatin is partitioned into distinct environments for each cell type (Figure 1.4d). It was initially proposed that LOCKs are relatively scarce in ES cells as few such chromatin domains could be detected. However, whether this reflects a true distinction in modification patterns between cell types, or a detection bias, has been questioned (Filion and van Steensel 2010). The nature of these compartments remains an area of active investigation as these structures could play a critical role in sequestering non-utilized regions of the genome, and thereby reducing the effective 'search-space' for gene regulatory machinery.

H3K27me3 blocks and Polycomb bodies.

Genomewide histone modification maps have also revealed large blocks of H3K27me3 in differentiated cells. An appreciation of these domains relied on new algorithms for identifying broad regions, rather than sharp peaks, of enrichment, as two recent studies illustrate. Pauler *et al.* used an algorithm called broad local enrichments (BLOCs) to identify H3K27me3 blocks that average 43 kb and overlap silent genes and intergenic regions (Pauler, Sloane et al. 2009). They find this pattern in numerous ChIP-chip and ChIP-seq datasets, and suggest that this is a common feature of H3K27me3 in differentiated cell types. The authors speculate that these H3K27me3 blocks may relate to Giemsa bands, as they observe alternating chromatin patterns along chromosomes. Hawkins *et al.* used ChromaBlocks to find similar H3K27me3 blocks in IMR90 fibroblasts, and further characterize their dynamics during differentiation (Hawkins, Hon et al.). This study suggests that these repressive domains are often seeded in ES cells and expand in differentiated cell types, apparently to confer cell type-specific repression (Figure 1.4d). As these domains have only recently been observed, little is known about their establishment or functional consequences. It is tempting to consider that, like H3K9me2 domains, H3K27me3 blocks mark distinct nuclear structures or regions. They potentially correspond to Polycomb bodies, discrete foci of silenced genes that have been observed by imaging and in situ hybridization in fly and human cells (Sexton, Schober et al. 2007). Although there is no data yet directly tying H3K27me3 blocks to these structures, there is indirect evidence linking the mark to compacted chromatin. H3K27me3 can promote recruitment of PRC1 (Simon and Kingston 2009). In turn, PRC1 may be required to maintain chromatin compaction at the Hox loci in ES cells (Eskeland, Leeb et al.). Together, these studies support connections between Polycomb regulation, histone

modifications and chromatin compartmentalization that promise to be an exciting area for further investigation.

Replication time zones.

In addition to delineating particular genomic elements, chromatin patterns gleaned through mapping studies also appear to relate to DNA replication timing (Figure 1.5). Rather than executing replication in a random fashion, the genome is divided into distinct replication time zones that average 1 Mb in size and tend to undergo DNA synthesis at coordinated times during S-phase (Goren and Cedar 2003). Plasmid injection experiments initially suggested a tight link between replication timing and histone H3 and H4 acetylation: regardless of sequence, a DNA fragment that is introduced into a cell in early S will be wrapped around acetylated histones, while the same fragment will be associated with deacetylated histones when injected in late S (Zhang, Xu et al. 2002). Genomewide profiling of replication timing in mouse and human cells revealed a correlation between replication domains and chromatin structure (Ryba, Hiratani et al. ; Karnani, Taylor et al. 2007). Early replicating zones associate with H3K4me1/2/3, H3K9 and H3K27 acetylation, and H3K36me3 and H4K20me1, while late replicating zones mostly correlate with H3K9me2, and to a lesser degree with H3K9me3 (Ryba, Hiratani et al.). More than simply correlative, subsequent studies have shown that the histone acetylation patterns directly influence the timing at which origins initiate replication (‘fire’) during S-phase in both yeast and mouse models (Vogelauer, Rubbi et al. 2002; Goren, Tabib et al. 2008). Of note, bivalent chromatin replicates early, despite being transcriptionally inactive, potentially reflecting an accessible and poised character (Azuara, Perry et al. 2006). Notably, boundaries between

replicating zones are also associated with a signature modification pattern – namely, peaks of H3K4me1/2/3, H3K27ac and H3K36me3. It has been speculated that the 'active' histone modifications might serve as functional boundary elements that block spreading of late-replicating heterochromatin. Together, the studies presented above illustrate global correspondences between histone modification patterns, replication timing and higher-order nuclear structures (Figure 1.5).

PERSPECTIVES AND FUTURE CHALLENGES

The implications of the growing panel of genomewide histone modification maps are several. At the level of the primary chromatin structure, the data suggest that histone modifications are indicative of functional genomic elements, gene expression, splicing patterns, and modes of repression. Together with perturbation studies on the mechanisms that write and read these marks, this insight may enable us to better understand and predict how normal or diseased cell types utilize and regulate their genomes. Additionally, these maps promote an appreciation of the three-dimensional organization of the genome. During the last few years, more and more pieces of the nuclear architecture 'jigsaw-puzzle' have been revealed. As we have discussed, histone modifications are intimately tied to large-scale repressive domains like LADs and Polycomb bodies, and broad patterns of replication time zones. Together with ongoing studies of additional structures such as transcription factories and nucleolus-associated domains (Nemeth, Conesa et al. ; Schoenfelder, Sexton et al. 2010), these findings are building a better understanding of the architecture of chromatin within the nucleus.

Several recent technological advances provide direction towards a molecular understanding of the spatial organization of chromatin. Lieberman-Aiden *et al.* scaled the chromosome conformation capture (3C) assay for unbiased genomewide identification of chromatin interactions (Hi-C) (Lieberman-Aiden, van Berkum et al. 2009). This approach revealed distinct spatial compartments distinguished by their degree of openness, but was limited in terms of the resolution with which it could distinguish interactions and compartmentalization. Fullwood *et al.* scaled the technology of a related approach that also incorporates an immunoprecipitation step (ChIA-PET) (Fullwood, Liu et al. 2009). They focused on the interaction network bound by oestrogen receptor α, and note numerous cases of chromatin looping for coordinated transcriptional regulation. Another important area of technology development relates to miniaturization and increasing the sensitivity of the assays so they may be compatible with small samples or even individual cells (Adli, Zhu et al. 2010; Goren, Ozsolak et al. 2010). High-resolution imaging approaches may also be instrumental in this regard. Combined with more powerful and integrative computational algorithms, such tools should ultimately enable every genomic region within a living cell to be tracked across differentiation, development, and disease.

Despite our increasing knowledge on various aspects of chromatin structure, we are still far from understanding the determinants of this structure. Relatively little is known about the complexes that introduce and maintain histone modification patterns. Even less is known about the way specific modification signatures, or 'states,' are read. How combinatorial options of chromatin 'writer' and 'reader' proteins facilitate more sophisticated and robust regulation of gene expression and genome function remains a key area of investigation. Detailed knowledge of global chromatin architecture along with these regulators represents a critical step towards

understanding how genetic, epigenetic, and environmental/stochastic factors drive context-specific genome regulation.

This era is an exciting time in biology, in which new genomic tools are validating or refuting dogmas developed through gene-specific analysis, as well as illuminating entirely unexpected principles. The pace of change is accelerating thanks to remarkable advances in DNA sequencing, increasing availability of epigenomic data in the public domain from NIH and international projects (Birney, Stamatoyannopoulos et al. 2007; Bernstein, Stamatoyannopoulos et al. 2010; Satterlee, Schubeler et al. 2010), and the rapid dissemination of these technologies into individual research laboratories. By changing our focus from 'gene-centered' to 'genomewide', such approaches hold much promise to enhance our understanding of genome architecture and its consequences on gene regulation, genome stability, cell phenotype, and organismal physiology in both health and disease.

GLOSSARY

DNaseI-seq: A method that distinguishes open chromatin regions based on their hypersensitivity to DNase I digestion. Sequencing these genomic fragments can generate genomewide maps of chromatin accessibility.

FAIRE-seq: "Formaldehyde Assisted Isolation of Regulatory Elements" followed by sequencing exploits the solubility of open chromatin in the aqueous phase during phenol:chloroform extraction to generate genomewide maps of soluble chromatin.

Sono-seq: A technique that relies on the increased sonication efficiency of open crosslinked chromatin to identify regions of increased accessibility genomewide.

MNase-seq: A method that distinguishes nucleosome positioning based on the ability of nucleosomes to protect associated DNA from digestion by micrococcal nuclease. Protected fragments are sequenced to produce genomewide maps of nucleosome localization.

CATCH-IT: "Covalent Attachment of Tags to Capture Histones and Identify Turnover" is an assay for measuring nucleosome turnover kinetics genomewide by metabolically labeling histones and profiling labeled DNA by microarray.

ChIP-seq: A method for mapping the distribution of histone modifications and chromatin-associated proteins genomewide that relies on immunoprecipitation with antibodies to modified histones or other chromatin proteins. The enriched DNA is sequenced to create genomewide profiles.

Hidden Markov Model: An HMM is a statistical model in which internal states are not visible but the outputs of these states are, and these outputs can be used to infer the internal states. This model can be used to determine biologically-relevant states from ChIP-seq datasets.

CpG islands: Genomic regions that are enriched for CpG dinucleotides, often occurring near constitutively active promoters. Mammalian genomes are otherwise depleted of CpGs due to preferential deamination of methylated cytosines.

DamID: A method for mapping the distribution of chromatin-associated proteins by fusing a protein of interest with *E.coli* DNA adenine methyltransferase (Dam), which methylates adenines proximal to the protein's binding sites, thus circumventing the need for antibodies.

Giemsa bands: Also known as G-bands, a characteristic banding pattern obtained by treating chromosomes with Giemsa stain. The intensity of Giemsa staining is correlated with genomic features. For instance, dark Giemsa bands usually are AT rich, have low gene density, and have higher densities of repeat elements.

Polycomb bodies: Discrete nuclear foci containing Polycomb proteins and their silenced target genes. These have been observed in both *D. melanogaster* and human cells by *in situ* hybridization.

3C: "Chromosome Conformation Capture" is a method to map chromosome interactions locally. It relies on increased frequency of intramolecular ligation between fragments in close three-dimensional proximity within the nucleus.

ACKNOWLEDGEMENTS

We thank Eric Mendenhall, Manching Ku, Richard Koche, and Esther Rheinbay for critical reading of the manuscript. We also thank members of the Bernstein Laboratory for insightful discussions. V.W.Z. was supported by a National Defense Science and Engineering Graduate Fellowship and a National Science Foundation Graduate Research Fellowship. A.G. was supported by an EMBO long term postdoctoral fellowship. B.E.B. is a Charles E. Culpeper Medical Scholar and Early Career Scientist of the Howard Hughes Medical Institute. Research in the Bernstein Laboratory is supported by funds from the Burroughs Wellcome Fund, HHMI, and the NIH.

REFERENCES

Adli, M., J. Zhu, et al. "Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors." Nat Methods **7**(8): 615-618.

Adli, M., J. Zhu, et al. (2010). "Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors." Nat Methods **7**(8): 615-618.

Andersson, R., S. Enroth, et al. (2009). "Nucleosomes are well positioned in exons and carry characteristic histone modifications." Genome Res **19**(10): 1732-1741.

Auerbach, R. K., G. Euskirchen, et al. (2009). "Mapping accessible chromatin regions using Sono-Seq." Proc Natl Acad Sci U S A **106**(35): 14926-14931.

Azuara, V., P. Perry, et al. (2006). "Chromatin signatures of pluripotent cell lines." Nat Cell Biol **8**(5): 532-538.

Barski, A., S. Cuddapah, et al. (2007). "High-resolution profiling of histone methylations in the human genome." Cell **129**(4): 823-837.

Bernstein, B. E., M. Kamal, et al. (2005). "Genomic maps and comparative analysis of histone modifications in human and mouse." Cell **120**(2): 169-181.

Bernstein, B. E., A. Meissner, et al. (2007). "The mammalian epigenome." Cell **128**(4): 669-681.

Bernstein, B. E., T. S. Mikkelsen, et al. (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." Cell **125**(2): 315-326.

Bernstein, B. E., J. A. Stamatoyannopoulos, et al. (2010). "The NIH Roadmap Epigenomics Mapping Consortium." Nat Biotechnol **28**(10): 1045-1048.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Blackledge, N. P., J. C. Zhou, et al. (2010). "CpG islands recruit a histone H3 lysine 36

demethylase." Mol Cell **38**(2): 179-190.

Boyer, L. A., K. Plath, et al. (2006). "Polycomb complexes repress developmental regulators in murine embryonic stem cells." Nature **441**(7091): 349-353.

Boyle, A. P., S. Davis, et al. (2008). "High-resolution mapping and characterization of open chromatin across the genome." Cell **132**(2): 311-322.

Coolen, M. W., C. Stirzaker, et al. "Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity." Nat Cell Biol **12**(3): 235-246.

Cui, K., C. Zang, et al. (2009). "Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation." Cell Stem Cell **4**(1): 80-93.

De Santa, F., I. Barozzi, et al. (2010). "A large fraction of extragenic RNA pol II transcription sites overlap enhancers." PLoS Biol **8**(5): e1000384.

Deal, R. B., J. G. Henikoff, et al. "Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones." Science **328**(5982): 1161-1164.

Dion, M. F., T. Kaplan, et al. (2007). "Dynamics of replication-independent histone turnover in budding yeast." Science **315**(5817): 1405-1408.

Down, T. A., V. K. Rakyan, et al. (2008). "A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis." Nat Biotechnol **26**(7): 779-785.

Durrin, L. K., R. K. Mann, et al. (1991). "Yeast histone H4 N-terminal sequence is required for promoter activation in vivo." Cell **65**(6): 1023-1031.

Ernst, J. and M. Kellis (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nat Biotechnol **28**(8): 817-825.

Eskeland, R., M. Leeb, et al. "Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination." Mol Cell **38**(3): 452-464.

Felsenfeld, G. and M. Groudine (2003). "Controlling the double helix." Nature **421**(6921): 448-453.

Filion, G. J., J. G. van Bemmel, et al. (2010). "Systematic protein location mapping reveals five principal chromatin types in Drosophila cells." Cell **143**(2): 212-224.

Filion, G. J. and B. van Steensel (2010). "Reassessing the abundance of H3K9me2 chromatin domains in embryonic stem cells." Nat Genet **42**(1): 4; author reply 5-6.

Finlan, L. E., D. Sproul, et al. (2008). "Recruitment to the nuclear periphery can alter expression of genes in human cells." PLoS Genet **4**(3): e1000039.

Fouse, S. D., Y. Shen, et al. (2008). "Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation." Cell Stem Cell **2**(2): 160-169.

Fullwood, M. J., M. H. Liu, et al. (2009). "An oestrogen-receptor-alpha-bound human chromatin interactome." Nature **462**(7269): 58-64.

Gilmour, D. S. and J. T. Lis (1984). "Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes." Proc Natl Acad Sci U S A **81**(14): 4275-4279.

Giresi, P. G., J. Kim, et al. (2007). "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin." Genome Res **17**(6): 877-885.

Goldberg, A. D., L. A. Banaszynski, et al. "Distinct factors control histone variant H3.3 localization at specific genomic regions." Cell **140**(5): 678-691.

Goren, A. and H. Cedar (2003). "Replicating by the clock." Nat Rev Mol Cell Biol **4**(1): 25-32.

Goren, A., F. Ozsolak, et al. (2010). "Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA." Nat Methods **7**(1): 47-49.

Goren, A., A. Tabib, et al. (2008). "DNA replication timing of the human beta-globin domain is controlled by histone modification at the origin." Genes Dev **22**(10): 1319-1324.

Guelen, L., L. Pagie, et al. (2008). "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions." Nature **453**(7197): 948-951.

Guenther, M. G., S. S. Levine, et al. (2007). "A chromatin landmark and transcription initiation

at most promoters in human cells." Cell **130**(1): 77-88.

Hansen, K. H., A. P. Bracken, et al. (2008). "A model for transmission of the H3K27me3 epigenetic mark." Nat Cell Biol **10**(11): 1291-1300.

Hargreaves, D. C., T. Horng, et al. (2009). "Control of inducible gene expression by signal-dependent transcriptional elongation." Cell **138**(1): 129-145.

Hawkins, R. D., G. C. Hon, et al. "Distinct epigenomic landscapes of pluripotent and lineage-committed human cells." Cell Stem Cell **6**(5): 479-491.

Hawkins, R. D., G. C. Hon, et al. "Next-generation genomics: an integrative approach." Nat Rev Genet **11**(7): 476-486.

Heintzman, N. D., G. C. Hon, et al. (2009). "Histone modifications at human enhancers reflect global cell-type-specific gene expression." Nature **459**(7243): 108-112.

Heintzman, N. D., R. K. Stuart, et al. (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." Nat Genet **39**(3): 311-318.

Hesselberth, J. R., X. Chen, et al. (2009). "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting." Nat Methods **6**(4): 283-289.

Hon, G., W. Wang, et al. (2009). "Discovery and annotation of functional chromatin signatures in the human genome." PLoS Comput Biol **5**(11): e1000566.

Irizarry, R. A., C. Ladd-Acosta, et al. (2009). "The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores." Nat Genet **41**(2): 178-186.

Ji, H., L. I. Ehrlich, et al. "Comprehensive methylome map of lineage commitment from haematopoietic progenitors." Nature.

Jirtle, R. L. and M. K. Skinner (2007). "Environmental epigenomics and disease susceptibility." Nat Rev Genet **8**(4): 253-262.

Jones, P. A. and S. B. Baylin (2007). "The epigenomics of cancer." Cell **128**(4): 683-692.

Kanhere, A., K. Viiri, et al. "Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2." Mol Cell **38**(5): 675-688.

Kaplan, N., I. K. Moore, et al. (2009). "The DNA-encoded nucleosome organization of a eukaryotic genome." Nature **458**(7236): 362-366.

Karnani, N., C. Taylor, et al. (2007). "Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas." Genome Res **17**(6): 865-876.

Kim, H., K. Kang, et al. (2009). "AEBP2 as a potential targeting protein for Polycomb Repression Complex PRC2." Nucleic Acids Res **37**(9): 2940-2950.

Kim, T. G., J. C. Kraus, et al. (2003). "JUMONJI, a critical factor for cardiac development, functions as a transcriptional repressor." J Biol Chem **278**(43): 42247-42255.

Kim, T. H., Z. K. Abdullaev, et al. (2007). "Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome." Cell **128**(6): 1231-1245.

Kim, T. H., L. O. Barrera, et al. (2005). "A high-resolution map of active promoters in the human genome." Nature **436**(7052): 876-880.

Kim, T. K., M. Hemberg, et al. "Widespread transcription at neuronal activity-regulated enhancers." Nature **465**(7295): 182-187.

Kolasinska-Zwierz, P., T. Down, et al. (2009). "Differential chromatin marking of introns and expressed exons by H3K36me3." Nat Genet **41**(3): 376-381.

Kornblihtt, A. R., I. E. Schor, et al. (2009). "When chromatin meets splicing." Nat Struct Mol Biol **16**(9): 902-903.

Ku, M., R. P. Koche, et al. (2008). "Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains." PLoS Genet **4**(10): e1000242.

Kumaran, R. I. and D. L. Spector (2008). "A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence." J Cell Biol **180**(1): 51-65.

Law, J. A. and S. E. Jacobsen "Establishing, maintaining and modifying DNA methylation patterns in plants and animals." Nat Rev Genet **11**(3): 204-220.

Lee, J. H. and D. G. Skalnik (2005). "CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex." J Biol Chem **280**(50): 41725-41731.

Lee, T. I., R. G. Jenner, et al. (2006). "Control of developmental regulators by Polycomb in human embryonic stem cells." Cell **125**(2): 301-313.

Li, B., M. Carey, et al. (2007). "The role of chromatin during transcription." Cell **128**(4): 707-719.

Li, G., R. Margueron, et al. "Jarid2 and PRC2, partners in regulating gene expression." Genes Dev **24**(4): 368-380.

Lieberman-Aiden, E., N. L. van Berkum, et al. (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." Science **326**(5950): 289-293.

Lister, R., M. Pelizzola, et al. (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences." Nature **462**(7271): 315-322.

Luco, R. F., Q. Pan, et al. "Regulation of alternative splicing by histone modifications." Science **327**(5968): 996-1000.

Margueron, R., N. Justin, et al. (2009). "Role of the polycomb protein EED in the propagation of repressive histone marks." Nature **461**(7265): 762-767.

Margueron, R. and D. Reinberg "Chromatin structure and the inheritance of epigenetic information." Nat Rev Genet **11**(4): 285-296.

Meissner, A., T. S. Mikkelsen, et al. (2008). "Genome-scale DNA methylation maps of pluripotent and differentiated cells." Nature **454**(7205): 766-770.

Mendenhall, E. M., R. P. Koche, et al. (2010). "GC-rich sequence elements recruit PRC2 in mammalian ES cells." PLoS Genet **6**(12): e1001244.

Mikkelsen, T. S., J. Hanna, et al. (2008). "Dissecting direct reprogramming through integrative genomic analysis." Nature **454**(7200): 49-55.

Mikkelsen, T. S., M. Ku, et al. (2007). "Genome-wide maps of chromatin state in pluripotent and

lineage-committed cells." Nature **448**(7153): 553-560.

Mohn, F., M. Weber, et al. (2008). "Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors." Mol Cell **30**(6): 755-766.

Nemeth, A., A. Conesa, et al. "Initial genomics of the human nucleolus." PLoS Genet **6**(3): e1000889.

Nilsen, T. W. and B. R. Graveley "Expansion of the eukaryotic proteome by alternative splicing." Nature **463**(7280): 457-463.

Ooi, S. K., C. Qiu, et al. (2007). "DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA." Nature **448**(7154): 714-717.

Orford, K., P. Kharchenko, et al. (2008). "Differential H3K4 methylation identifies developmentally poised hematopoietic genes." Dev Cell **14**(5): 798-809.

Pan, G., S. Tian, et al. (2007). "Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells." Cell Stem Cell **1**(3): 299-312.

Park, P. J. (2009). "ChIP-seq: advantages and challenges of a maturing technology." Nat Rev Genet **10**(10): 669-680.

Pasini, D., P. A. Cloos, et al. "JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells." Nature **464**(7286): 306-310.

Pauler, F. M., M. A. Sloane, et al. (2009). "H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome." Genome Res **19**(2): 221-233.

Peng, J. C., A. Valouev, et al. (2009). "Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells." Cell **139**(7): 1290-1302.

Phillips, J. E. and V. G. Corces (2009). "CTCF: master weaver of the genome." Cell **137**(7): 1194-1211.

Rahl, P. B., C. Y. Lin, et al. "c-Myc regulates transcriptional pause release." Cell **141**(3): 432-445.

Ramirez-Carrozzi, V. R., D. Braas, et al. (2009). "A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling." Cell **138**(1): 114-128.

Reddy, K. L., J. M. Zullo, et al. (2008). "Transcriptional repression mediated by repositioning of genes to the nuclear lamina." Nature **452**(7184): 243-247.

Ryba, T., I. Hiratani, et al. "Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types." Genome Res **20**(6): 761-770.

Satterlee, J. S., D. Schubeler, et al. (2010). "Tackling the epigenome: challenges and opportunities for collaboration." Nat Biotechnol **28**(10): 1039-1044.

Schoenfelder, S., T. Sexton, et al. (2010). "Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells." Nat Genet **42**(1): 53-61.

Schones, D. E., K. Cui, et al. (2008). "Dynamic regulation of nucleosome positioning in the human genome." Cell **132**(5): 887-898.

Schones, D. E. and K. Zhao (2008). "Genome-wide approaches to studying chromatin modifications." Nat Rev Genet **9**(3): 179-191.

Schwartz, S., E. Meshorer, et al. (2009). "Chromatin organization marks exon-intron structure." Nat Struct Mol Biol **16**(9): 990-995.

Seila, A. C., J. M. Calabrese, et al. (2008). "Divergent transcription from active promoters." Science **322**(5909): 1849-1851.

Sexton, T., H. Schober, et al. (2007). "Gene regulation through nuclear organization." Nat Struct Mol Biol **14**(11): 1049-1055.

Shen, X., W. Kim, et al. (2009). "Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells." Cell **139**(7): 1303-1314.

Shilatifard, A. (2008). "Molecular implementation and physiological roles for histone H3 lysine 4 (H3K4) methylation." Curr Opin Cell Biol **20**(3): 341-348.

Simon, J. A. and R. E. Kingston (2009). "Mechanisms of polycomb gene silencing: knowns and

unknowns." Nat Rev Mol Cell Biol **10**(10): 697-708.

Sing, A., D. Pannell, et al. (2009). "A vertebrate Polycomb response element governs segmentation of the posterior hindbrain." Cell **138**(5): 885-897.

Solomon, M. J. and A. Varshavsky (1985). "Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures." Proc Natl Acad Sci U S A **82**(19): 6470-6474.

Stock, J. K., S. Giadrossi, et al. (2007). "Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells." Nat Cell Biol **9**(12): 1428-1435.

Straussman, R., D. Nejman, et al. (2009). "Developmental programming of CpG island methylation profiles in the human genome." Nat Struct Mol Biol **16**(5): 564-571.

Thomson, J. P., P. J. Skene, et al. "CpG islands influence chromatin structure via the CpG-binding protein Cfp1." Nature **464**(7291): 1082-1086.

Tilgner, H., C. Nikolaou, et al. (2009). "Nucleosome positioning as a determinant of exon recognition." Nat Struct Mol Biol **16**(9): 996-1001.

Tsai, M. C., O. Manor, et al. "Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes." Science.

Visel, A., M. J. Blow, et al. (2009). "ChIP-seq accurately predicts tissue-specific activity of enhancers." Nature **457**(7231): 854-858.

Visel, A., E. M. Rubin, et al. (2009). "Genomic views of distant-acting enhancers." Nature **461**(7261): 199-205.

Vogelauer, M., L. Rubbi, et al. (2002). "Histone acetylation regulates the time of replication origin firing." Mol Cell **10**(5): 1223-1233.

Wang, Z., C. Zang, et al. (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome." Nat Genet **40**(7): 897-903.

Weber, M., I. Hellmann, et al. (2007). "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome." Nat Genet **39**(4): 457-466.

Wen, B., H. Wu, et al. (2009). "Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells." Nat Genet **41**(2): 246-250.

Wendt, K. S., K. Yoshida, et al. (2008). "Cohesin mediates transcriptional insulation by CCCTC-binding factor." Nature **451**(7180): 796-801.

Woo, C. J., P. V. Kharchenko, et al. "A region of the human HOXD cluster that confers polycomb-group responsiveness." Cell **140**(1): 99-110.

Zhang, J., F. Xu, et al. (2002). "Establishment of transcriptional competence in early and late S phase." Nature **420**(6912): 198-202.

Zhao, J., B. K. Sun, et al. (2008). "Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome." Science **322**(5902): 750-756.

Zhao, X. D., X. Han, et al. (2007). "Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells." Cell Stem Cell **1**(3): 286-298.

Zilberman, D., D. Coleman-Derr, et al. (2008). "Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks." Nature **456**(7218): 125-129.

Chapter 2:

GC-rich Sequence Elements, rather than YY1,

Recruit PRC2 in Mammalian ES Cells

GC-rich Sequence Elements, rather than YY1, Recruit PRC2 in Mammalian ES Cells

This Chapter was revised from the following publication to reflect the contributions of Vicky W. Zhou:

"GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells." (2010)

PLoS Genetics 6(12): e1001244

Eric M. Mendenhall[1,2,3], *Richard P. Koche**[1,2,3,4], *Thanh Truong*[1,2,3], *Vicky W. Zhou*[1,2,3,5], *Biju Issac*[1,2,3], *Andrew S. Chi*[1,2,3,6], *Manching Ku*[1,2,3], *Bradley E. Bernstein*[1,2,3]

1. Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA.

2. Center for Systems Biology and Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA

3. Broad Institute of Harvard and MIT, Cambridge, MA, USA.

4. Division of Health Sciences and Technology, MIT, Cambridge, MA, USA

5. Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA

6. Neuro-Oncology Division, Department of Neurology, Massachusetts General Hospital, Boston, MA, USA.

* Equal Contributions

Correspondence and requests for materials should be addressed to B.E.B. (Bernstein.Bradley@mgh.harvard.edu).

AUTHOR CONTRIBUTIONS

V.W.Z. and B.E.B. conceived and designed the YY1 ChIP and ChIP-seq experiments. V.W.Z. performed the YY1 experiments in Figures 2.3D, 2.4C, and 2.4D. V.W.Z., E.M., and B.E.B. conceived and designed the 22KB and –CGI ChIP experiments. V.W.Z. performed the 22KB, –CGI, and Gene Desert CGI ChIP experiments in Figures 2.2A, 2.2B, and 2.3C.

R.P.K., E.M.M., and B.E.B. developed the hypothesis for CpG islands as causal Polycomb recruitment elements. E.M.M., R.P.K., and B.E.B. conceived and designed the overall strategy for testing Polycomb recruitment with a transgenic chromatin assay, as shown in Figure 2.1. E.M.M. performed the 44KB and Gene Desert ChIP experiments in Figures 2.2A, 2.2B, and 2.3B. R.P.K. performed the %GC and motif analysis in Figures 2.4A and 2.4B. M.K. performed the Ring1B ChIP experiments in Figures 2.2B and 2.3C.

ABSTRACT

Polycomb proteins are epigenetic regulators that localize to developmental loci in the early embryo where they mediate lineage-specific gene repression. In *Drosophila*, these repressors are recruited to sequence elements by DNA binding proteins associated with Polycomb repressive complex 2 (PRC2). However, the sequences that recruit PRC2 in mammalian cells have remained obscure. To address this, we integrated a series of engineered bacterial artificial chromosomes into embryonic stem (ES) cells and examined their chromatin. We found that a 44 kb region corresponding to the Zfpm2 locus initiates *de novo* recruitment of PRC2. We then pinpointed a CpG island within this locus as both necessary and sufficient for PRC2 recruitment. We found that YY1 binding was not necessary for this recruitment, and that genomewide YY1 localization does not correlate with PRC1 or PRC2 localization. Our findings demonstrate a causal role for GC-rich sequences in PRC2 recruitment, and suggest that YY1 is not directly involved in PRC2 recruitment in mammalian genomes.

AUTHOR SUMMARY

Key developmental genes are precisely turned on or off during development, thus creating a complex, multi-tissue embryo. The mechanism that keeps genes off, or repressed, is crucial to proper development. In embryonic stem cells, Polycomb repressive complex 2 (PRC2) is recruited to the promoters of these developmental genes, and helps to maintain repression in the appropriate tissues through development. How PRC2 is initially recruited to these genes in the early embryo remains elusive. Here we experimentally demonstrate that stretches of GC-rich

DNA, termed CpG islands, can initiate recruitment of PRC2 in embryonic stem cells. This supports a model where inactive GC-rich DNA can itself suffice to recruit PRC2 even in the absence of more complex DNA sequence motifs, such as YY1.

INTRODUCTION

Polycomb proteins are epigenetic regulators required for proper gene expression patterning in metazoans. The proteins reside in two main complexes, termed Polycomb repressive complex 1 and 2 (PRC1 and PRC2). PRC2 catalyzes histone H3 lysine 27 tri-methylation (K27me3), while PRC1 catalyzes histone H2A ubiquitination and mediates chromatin compaction (Schwartz and Pirrotta 2007; Schuettengruber, Ganapathi et al. 2009). PRC1 and PRC2 are initially recruited to target loci in the early embryo where they subsequently mediate lineage-specific gene repression. In embryonic stem (ES) cells, the complexes localize to thousands of genomic sites, including many developmental loci (Boyer, Plath et al. 2006; Lee, Jenner et al. 2006; Ku, Koche et al. 2008). These target loci are not yet stably repressed, but instead maintain a "bivalent" chromatin state, with their chromatin enriched for the activating histone mark, H3 lysine 4 tri-methylation (K4me3), together with the repressive K27me3 (Azuara, Perry et al. 2006; Bernstein, Mikkelsen et al. 2006). In the absence of transcriptional induction, PRC1 and PRC2 remain at target loci and mediate repression through differentiation. The mechanisms that underlie stable association of the complexes remain poorly understood, but likely involve interactions with the modified histones (Cao, Wang et al. 2002; Czermin, Melfi et al. 2002; Kuzmichev, Nishioka et al. 2002; Hansen, Bracken et al. 2008; Margueron, Justin et al. 2009).

Proper localization of PRC1 and PRC2 in the pluripotent genome is central to the complex developmental regulation orchestrated by these factors. However, the sequence determinants that underlie this initial landscape remain obscure. Polycomb recruitment is best understood in *Drosophila,* where sequence elements termed Polycomb response elements (PREs) are able to direct these repressors to exogenous locations (Ringrose and Paro 2007). PREs contain clusters of motifs recognized by DNA binding proteins such as Pho, Zeste and GAGA, which in turn recruit PRC2 (Simon, Chiang et al. 1993; Wang, Brown et al. 2004; Dejardin, Rappailles et al. 2005; Tolhuis, de Wit et al. 2006). Despite extensive study, neither PRE sequence motifs nor binding profiles of PRC2-associated DNA binding proteins are sufficient to fully predict PRC2 localization in the *Drosophila* genome (Negre, Hennetin et al. 2006; Schwartz, Kahn et al. 2006; Tolhuis, de Wit et al. 2006; Schuettengruber, Ganapathi et al. 2009).

While protein homologs of PRC1 and PRC2 are conserved in mammals, DNA sequence homologs of *Drosophila* PREs appear to be lacking in mammalian genomes (Ringrose and Paro 2007). Moreover, it remains controversial whether the DNA binding proteins associated with PRC2 in *Drosophila* have functional homologs in mammals. The most compelling candidate has been YY1, a Pho homolog that rescues gene silencing when introduced into Pho-deficient *Drosophila* embryos (Atchison, Ghias et al. 2003). YY1 has been implicated in PRC2-dependent silencing of tumor suppressor genes in human cancer cells (Ko, Hsu et al. 2008). However, this transcription factor has also been linked to numerous other functions, including imprinting, DNA methylation, B-cell development and ribosomal protein gene transcription (Sui, Affar el et al. 2004; Liu, Schmidt-Supprian et al. 2007; Xi, Yu et al. 2007; Kim, Kang et al. 2009; Yue, Kang et al. 2009).

Recently, researchers identified two DNA sequence elements able to confer Polycomb

repression in mammalian cells. Sing and colleagues identified a murine PRE-like element that regulates the MafB gene during neural development (Sing, Pannell et al. 2009). These investigators defined a critical 1.5 kb sequence element that is able to recruit PRC1, but not PRC2 in a transgenic cell assay. Woo and colleagues identified a 1.8 kb region of the human HoxD cluster that recruits both PRC1 and PRC2 and represses a reporter construct in mesenchymal tissues (Woo, Kharchenko et al.). Both groups note that their respective PRE regions contain YY1 motifs. Mutation of the YY1 sites in the HoxD PRE resulted in loss of PRC1 binding and partial loss of repression, while comparatively, deletion of a separate highly conserved region from this element completely abrogated PRC1 and PRC2 binding as well as repression (Woo, Kharchenko et al.).

In addition to these locus-specific investigations, genomic studies have sought to define PRC2 targets and determinants in a systematic fashion. The Ezh2 and Suz12 subunits have been mapped in mouse and human ES cells by chromatin immunoprecipitation and microarrays (ChIP-chip) or high-throughput sequencing (ChIP-Seq) (Boyer, Plath et al. 2006; Bracken, Dietrich et al. 2006; Lee, Jenner et al. 2006; Ku, Koche et al. 2008). Such studies have highlighted global correlations between PRC2 targets and CpG islands (Ku, Koche et al. 2008; Mohn, Weber et al. 2008) as well as highly-conserved genomic loci (Bernstein, Mikkelsen et al. 2006; Lee, Jenner et al. 2006; Tanay, O'Donnell et al. 2007). Recently, Jarid2 has been shown to associate with PRC2 and to be required for proper genomewide localization of the complex (Li, Margueron et al. ; Pasini, Cloos et al. ; Peng, Valouev et al. 2009; Shen, Kim et al. 2009). Intriguingly, Jarid2 contains an ARID and a Zinc-finger DNA-binding domain. However, it is unclear how Jarid2 could account for PRC2 targeting given the lack of sequence specificity and the low affinity of its DNA binding domains (Li, Margueron et al. ; Kim, Kraus et al. 2003). In

summary, a variety of sequence elements including CpG islands, conserved elements and YY1 motifs have been implicated in Polycomb targeting in mammalian cells. Causality has only been demonstrated in two specific instances and a unifying view of the determinants of Polycomb recruitment remains elusive.

Here we present the identification of multiple sequence elements capable of recruiting PRC2 in mammalian ES cells. This was achieved through an experimental approach in which engineered bacterial artificial chromosomes (BACs) were stably integrated into the ES cell genome. Evaluation of a series of modified BACs specifically identified a 1.7 kb DNA fragment that is both necessary and sufficient for PRC2 recruitment. The fragment does not share sequence characteristics of *Drosophila* PREs and lacks YY1 binding sites, but rather corresponds to an annotated CpG island. Based on this result and a genomewide analysis of PRC2 target sequences, we propose that GC-rich sequence elements, rather than YY1, play causal roles in the initial localization of PRC2 and the subsequent coordination of epigenetic controls during mammalian development.

RESULTS

Recruitment of Polycomb repressors to a Bacterial Artificial Chromosome Integrated into ES Cells

To identify DNA sequences capable of recruiting Polycomb repressors in mammalian cells, we engineered human BACs that correspond to genomic regions bound by these proteins in human ES cells (Figure 2.1).

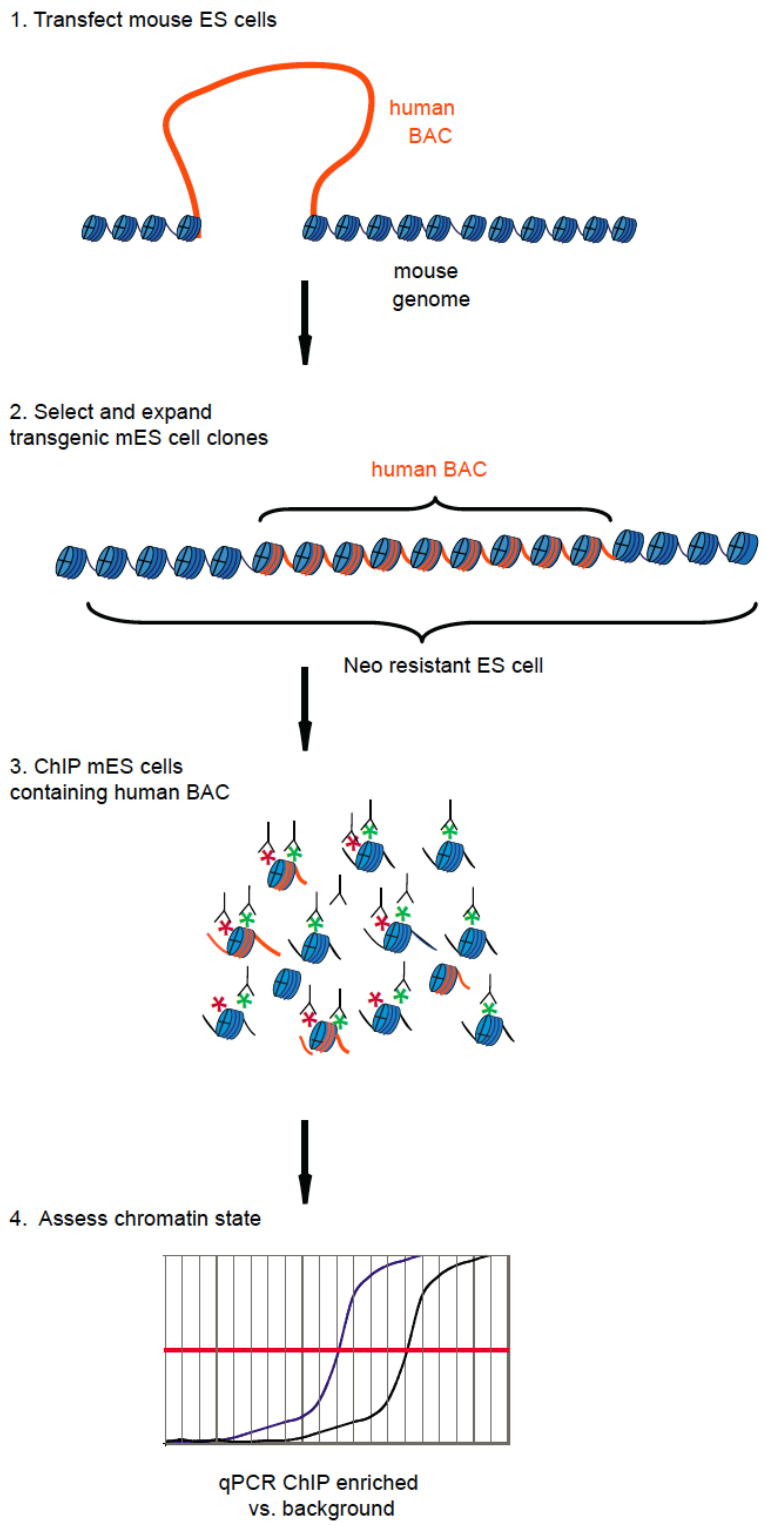

Figure 2.1. Schematic of Transgenic Chromatin Assay

A schematic of the transgenic chromatin assay that was used to examine the role of DNA sequence in determining histone modification patterns in embryonic stem cells.

We initially targeted a region of the human Zfpm2 (hZfpm2) locus, which encodes a developmental transcription factor involved in heart and gonad development (Tevosian, Albrecht et al. 2002). In ES cells, the endogenous locus recruits PRC1 and PRC2, and is enriched for the bivalent histone modifications, K4me3 and K27me3 (Figure 2.2A). We used recombineering to engineer a 44 kb BAC containing this locus and a neomycin selection marker. The modified BAC was electroporated into mouse ES cells, and individual transgenic ES cell colonies containing the full length BAC were expanded (Figure 2.1).

We used ChIP and quantitative PCR (ChIP-qPCR) with human specific primers to examine the chromatin state of the newly incorporated hZfpm2 locus. This analysis revealed strong enrichment for K27me3 and K4me3 (Figure 2.2B). In addition, we explicitly tested for direct binding of the Polycomb repressive complexes using antibody against the PRC1 subunit, Ring1B, or the PRC2 subunit, Ezh2. We detected robust enrichment for both complexes in the vicinity of the hZfpm2 gene promoter (Figure 2.2B). To confirm this result and eliminate the possibility of integration site effects, we tested two additional transgenic hZfpm2 ES cell clones with unique integration sites, and in both cases, we observed a bivalent chromatin state analogous to the endogenous loci. These results suggest that DNA sequence is sufficient to initiate *de novo* recruitment of Polycomb in ES cells.

Distinguishing Polycomb Recruiting Sequences in the Zfpm2 BAC

We next sought to define the sequences within the hZfpm2 BAC required for recruitment of Polycomb repressors. First, we re-engineered the 44 kb hZfpm2 BAC to remove 20 kb of flanking sequences that contained distal non-coding conserved sequence elements (Figure 2.2A).

Figure 2.2. Recruitment of Polycomb Repressors to a BAC Integrated into ES Cells

(A) ChIP-Seq tracks depict enrichment of K27me3 (the modification catalyzed by PRC2), Ezh2 (the enzymatic component of PRC2), and K4me3 across the endogenous hZfpm2 locus in human ES cells. Primers and constructs used in this study are indicated below the gene track. (B) BAC constructs from (A) containing the hZfpm2 locus were stably integrated into mouse ES cells. ChIP-qPCR enrichments are shown for K4me3, K27me3, Ezh2, and the PRC1 component Ring1b across the locus. The integrated locus adopts a 'bivalent' chromatin state with K27me3 and K4me3 in all constructs except the ΔCGI BAC. The locations of PCR amplicons are designated on the horizontal axis. Error bars show standard error of the mean (SEM) for n=3 (44 kb) or n=2 (22 kb; ΔCGI) biological replicates.

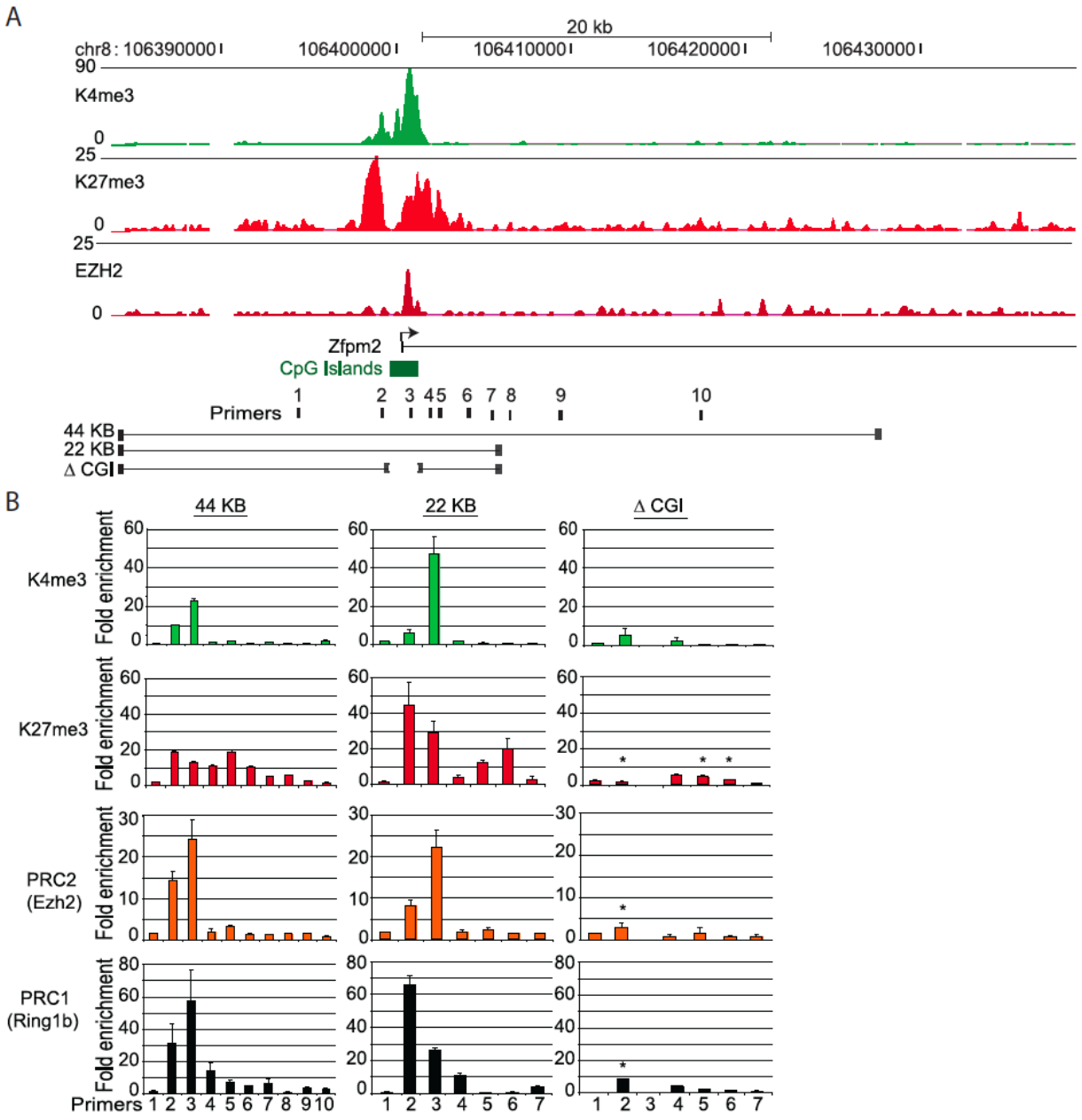

Figure 2.2 (Continued).

When we integrated the resulting 22 kb construct into ES cells we found that it robustly enriches for PRC1, PRC2, K4me3 and K27me3 (Figure 2.2B). Hence, these particular distal elements do not appear to be required for the recruitment of the complexes. Next, we considered the necessity of the CpG island which corresponds to the peak of Ezh2 enrichment in ChIP-Seq profiles (Figure 2.2A). We excised a 1.7 kb fragment containing the CpG island, and integrated the resulting BAC (ΔCGI) into ES cells. The ΔCGI BAC failed to recruit PRC1 or PRC2, and showed significantly reduced K27me3 levels relative to the other constructs (Figure 2.2B). This suggests that the CpG island is essential for recruitment of Polycomb proteins to the hZfpm2 locus.

A 1.7 kb CpG island is Sufficient to Recruit PRC2 to an Exogenous Locus

We next asked whether the hZfpm2 CpG island is sufficient to recruit Polycomb repressors to an exogenous locus. To test this, we selected an unremarkable gene desert region on human chromosome 1 that shows no enrichment for PRC1, PRC2 or K27me3 in ES Cells (Figure 2.3A). We also verified that the gene desert BAC alone does not show any enrichment for K27me3 or Ezh2 when integrated into ES cells (Figure 2.3B). Using recombineering, we inserted the 1.7 kb sequence that corresponds to the hZfpm2 CpG island into the gene desert BAC. The resulting construct was integrated into mouse ES cells and three independent clones were evaluated. ChIP-qPCR analysis revealed strong enrichment for K27me3, K4me3 and PRC2 over the inserted CpG island (Figure 2.3C). In contrast, we observed relatively little enrichment for the PRC1 subunit Ring1B (Figure 2.3C). We confirmed the specificity of these enrichments with primers that span the boundary between the insertion and adjacent gene desert sequence.

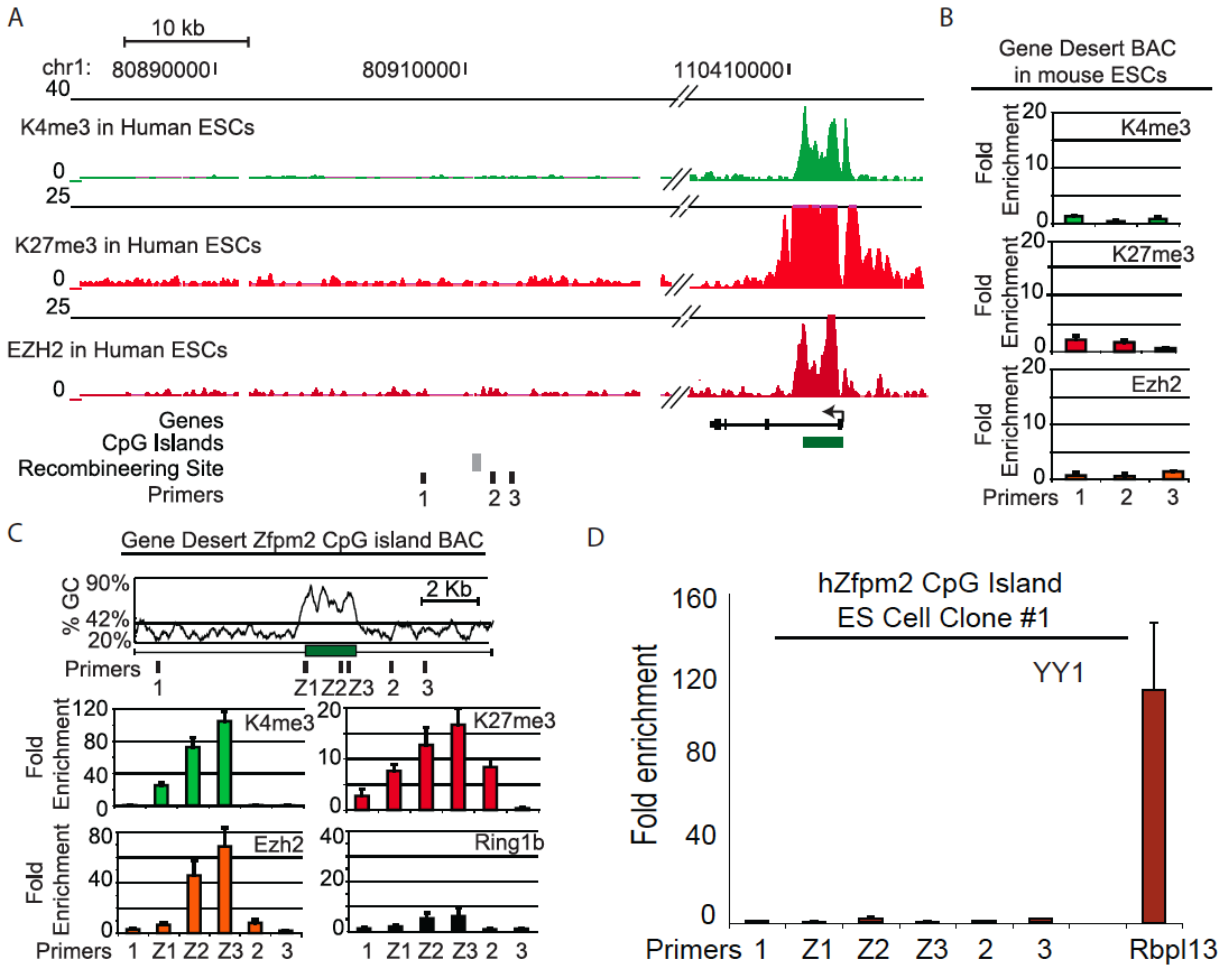

Figure 2.3. A 1.7 kb GC-rich Sequence Element is Sufficient to Recruit PRC2

(A) ChIP-Seq tracks show no enrichment for K4me3, K27me3 or Ezh2 in human ES cells across the gene desert region. For comparison a nearby locus is shown. The recombineering site and primers used are indicated below the tracks. (B) The gene desert BAC shows no enrichment of K4me3, K27me3 or PRC2 upon integration in mouse ES cells. (C) The hZfpm2 CpG island is depicted at the site of insertion into the gene desert BAC, along with the corresponding GC percentage (42% indicates genome average) and primers used for qPCR. Underlying plots represent ChIP-qPCR enrichment of K4me3, K27me3, PRC2 (Ezh2), and PRC1 (Ring1b) at the indicated sites (n=2 biological replicates). (D) The Zfpm2 Gene Desert BAC shows no enrichment of YY1, in contrast to the promoter of Rpl13a. Error bars equal to SEM (n = 2).

Notably, K27me3 enrichment was detected across the gene desert locus up to 2.5 kb from the inserted CpG island (Figure 2.3C). This indicates that the localized CpG island can initiate K27me3 that then spreads into adjacent sequence. Lastly we found no YY1 enrichment across the CpG island by ChIP-qPCR (Figure 2.3D). Together, these data suggest that the hZfpm2 CpG island contains the necessary signals for PRC2 recruitment but is insufficient to confer robust PRC1 association, and that YY1 binding is not necessary for PRC2 recruitment.

YY1 is not Directly Involved in PRC2 Recruitment in Mammalian ES Cells

The functionality of a CpG island in PRC2 recruitment is consistent with prior observations that a majority of PRC2 sites in ES cells correspond to CpG islands (Lee, Jenner et al. 2006; Ku, Koche et al. 2008). We therefore considered whether specific signals within the Zfpm2 CpG island might underlie its capacity to recruit PRC2.

First, we searched for sequence motifs analogous to the PREs that recruit PRC2 in *Drosophila.* We focused on motifs recognized by YY1, the nearest mammalian homolog of the *Drosophila* recruitment proteins. Notably, both of the recently described mammalian PREs contain YY1 motifs (Woo, Kharchenko et al. ; Sing, Pannell et al. 2009). The 44 kb hZfpm2 BAC contains 11 instances of the consensus YY1 motif. However, none of these reside within the CpG island (Figure 2.4A-B) (see Methods). We also examined YY1 binding directly in ES cells and NS cells using ChIP-Seq. Consistent with prior reports, YY1 binding is evident at the 5' ends of many highly expressed genes, including those encoding ribosomal proteins, and is also seen at the imprinted Peg3 locus (Figure 2.4C) (Kim, Kang et al. 2009). However, no YY1 enrichment is evident at the Zfpm2 locus.

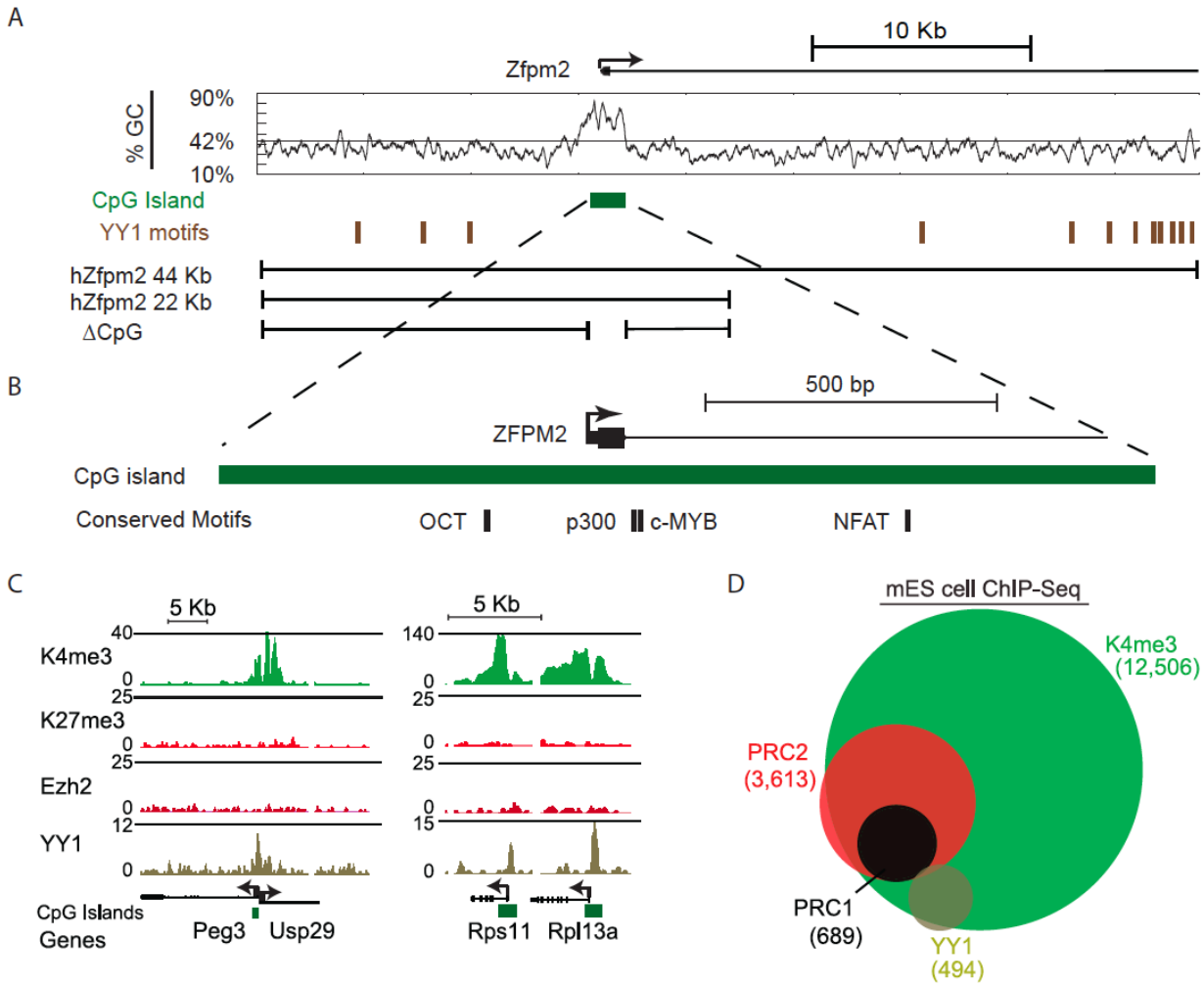

Figure 2.4. YY1 is not Directly Involved in PRC2 Recruitment in Mammalian ES Cells

(A) The GC-richness and locations of YY1 motifs for the Zfpm2 locus are shown. (B) The 1.7 kb CpG island contains 4 conserved motifs (see Methods). (C) ChIP-Seq was used to profile the mammalian Pho homolog YY1 in mouse ES cells. Genome browser views show ChIP-Seq enrichment signals for K4me3, K27me3, Ezh2 and YY1 for YY1 target loci. (D) Venn diagram shows overlap of K4me3, Ezh2, Ring1b, and YY1 at promoters in mES cells.

Moreover, at a global level, YY1 shows almost no overlap with PRC2 or PRC1, but instead co-localizes with genomic sites marked exclusively by K4me3 (Figure 2.4D). Thus, although YY1 may contribute to Polycomb-mediated repression through distal interactions or in *trans*, it does not appear to be directly involved in PRC2 recruitment in ES cells.

DISCUSSION

Several lines of evidence suggest that the initial landscape of Polycomb complex binding is critical for proper patterning of gene expression in metazoan development (Ringrose and Paro 2007; Schwartz and Pirrotta 2007; Schuettengruber, Ganapathi et al. 2009). Failure of these factors to engage their target loci in embryogenesis has been linked to a loss of epigenetic repression at later stages. Accordingly, the determinants that localize Polycomb complexes at the pluripotent stage are almost certainly essential to the global functions of these repressors through development.

We find that DNA sequence is sufficient for proper localization of Polycomb repressive complexes in ES cells, and specifically identify a CpG island within the Zfpm2 locus as being critical for recruitment. We provide evidence that YY1 is not directly involved in PRC2 recruitment in ES cells.

Several possible mechanistic models could explain the causality of GC-rich DNA elements in PRC2 recruitment. First, we note that CpG islands have been shown to destabilize nucleosomes in mammalian cells (Ramirez-Carrozzi, Braas et al. 2009). At transcriptionally inactive loci, this property could increase their accessibility to PRC2-associated proteins with DNA affinity but low sequence specificity, such as Jarid2 or AEBP2 (Li, Margueron et al. ;

Pasini, Cloos et al. ; Kim, Kang et al. 2009; Peng, Valouev et al. 2009; Shen, Kim et al. 2009). Although this association would be abrogated by transcriptional activity at most CpG islands, those lacking activation signals would remain permissive to PRC2 association. In support of this model, PRC2 targets in ES cells are also enriched for H2A.Z and H3.3, histone variants linked to nucleosome exchange dynamics (Goldberg, Banaszynski et al. ; Creyghton, Markoulaki et al. 2008). Alternatively or in addition, targeting could be supported by DNA binding proteins with affinity for low complexity GC-rich motifs or CpG dinucleotides, such as CXXC domain proteins (Tate, Lee et al. 2009). Localization may also be promoted or stabilized by long and short non-coding RNAs (Kanhere, Viiri et al. ; Tsai, Manor et al. ; Rinn, Kertesz et al. 2007; Zhao, Sun et al. 2008) as well as by the demonstrated affinity of PRC2 for its product, H3K27me3 (Hansen, Bracken et al. 2008; Margueron, Justin et al. 2009). Notably, PRC2 recruitment in ES cells appears distinct from that in *Drosophila*, as we do not find evidence for involvement of PRE-like sequence motifs or mammalian homologues such as YY1.

It should be emphasized that PRC2 localization does not necessarily equate with epigenetic repression. Indeed virtually all PRC2 bound sites in ES cells, and all CpG islands tested here, are also enriched for K4me3, and presumably poised for activation upon differentiation. Epigenetic repression during differentiation may require PRC1 and thus depend on additional binding determinants. YY1 remains an intriguing candidate in this regard, given prior evidence for physical and genetic interactions with PRC1 (Garcia, Marcos-Gutierrez et al. 1999; Lorente, Perez et al. 2006). YY1 consensus motifs are present in the Polycomb-dependent silencing elements recently identified in the MafB and HoxD loci. Interestingly, the HoxD element combines a CpG island with a cluster of conserved YY1 motifs. Mutation of the motifs abrogated PRC1 binding but left PRC2 binding intact. Still, the fact that only a small fraction of

documented PRC2 and PRC1 sites have YY1 motifs or binding suggests that this transcription factor may act indirectly and/or explain only a subset of cases. Nonetheless, it is likely that a fully functional epigenetic silencer would require a combination of features, including a GC-rich PRC2 element as well as appropriate elements to recruit PRC1. Further study is needed to expand the rules for PRC2 binding to include a global definition of PRC1 determinants and ultimately, to understand how the initial landscape facilitates the maintenance of gene expression programs in the developing organism.

Finally, we note that here, we used Ezh2 and Ring1B as proxies for the entire PRC2 and PRC1 complexes, respectively. While EZH2 and Ring1B are catalytic components and likely to be a member of every PRC2 or PRC1 complex, other members may only bind to a subset of targets. Attempts to characterize the binding of additional PRC2 and PRC1 components using ChIP-seq often failed due to the lack of high-quality antibodies and the difficulty of crosslinking loosely bound proteins. Thus, to fully understand the intricacies of PRC2 and PRC1 binding, new technology is needed to more reliably characterize chromatin proteins. We discuss one such technology in the next Chapter.

METHODS

BAC Construct Design

BAC constructs CTD331719L (‘Zfpm2 44’) and CTD-3219L19 (‘Gene Desert’) were obtained from Open Biosystems. Recombineering was done using the RedET system (Open Biosystems) in DH10B cells. Homology arms 200-500 bp in length were PCR amplified and

cloned into a PGK; Neomycin cassette (Gene Bridges). This cassette was used to recombineer all BACs to enable selection in mammalian cells. The 22 kb hZfpm2 BAC was created by restricting the hZfpm2 BAC at two sites using ClaI, and re-ligating the BAC lacking the intervening sequence. The CpG island was excised from the 22 kb hZfpm2 BAC by amplification of flanking homology arms, and cloned into a construct containing an adjacent ampicillin cassette (Frt-amp-Frt; Gene Bridges). After recombination, the ampicillin cassette was removed using Flp-recombinase and selection for clones that lost ampicillin resistance (Flp-706; Gene Bridges). PCR across the region confirmed excision of the CpG island. For the Gene Desert BAC, the Zfpm2 CpG island was amplified with primers containing XhoI sites and cloned into the Frt-amp-Frt vector that contains homology arms from the Gene Desert region. The final constructs were confirmed by sequencing across recombination junctions.

Transgenic ES Cell and ChIP Experiments

ES cells (V6.5) were maintained in ES cell medium (DMEM; Dulbecco's modified Eagle's medium) supplemented with 15% fetal calf serum (Hyclone), 0.1 mM ß-mercaptoethanol (Sigma), 2 mM Glutamax, 0.1 mM non-essential amino acid (NEAA; Gibco) and 1000U/ml recombinant leukemia inhibitory factor (ESGRO; Chemicon). Roughly 50 μg of linearized BAC was nucleofected using the mouse ES cell nucleofector kit (Lonza) into 10^6 mouse ES cells, and selected 7-10 days with 150 μg/ml Geneticin (Invitrogen) on Neomycin resistant MEFs (Millipore). Individual resistant colonies were picked, expanded and tested for integration of the full length BAC by PCR.

For each construct, between one and three ES cell clones were expanded and subjected to

ChIP using antibody against K4me3 (Abcam ab8580 or Upstate/Millipore 07-473), K27me3 (Upstate/Millipore 07-449), Ezh2 (Active Motif 39103 or 39639), or Ring1B (MBL International d139-3) as described previously (Bernstein, Mikkelsen et al. 2006; Mikkelsen, Ku et al. 2007; Ku, Koche et al. 2008). ChIP DNA was quantified by Quant-iT Picogreen dsDNA Assay Kit (Invitrogen). ChIP enrichments were assessed by quantitative PCR analysis on an ABI 7500 with 0.25 ng ChIP DNA and an equal mass of un-enriched input DNA. Enrichments were calculated from 2 or 3 biologically independent ChIP experiments. For K27me3, and Ezh2 enrichment, background was subtracted by normalizing over a negative genomic control. Error bars represent standard error of the mean (SEM). We confirmed that the human specific primers do not non-specifically amplify mouse genomic DNA.

Genomic and Computational Analysis

Genomewide maps of YY1 binding sites were determined by ChIP-Seq as described previously (Mikkelsen, Ku et al. 2007). Briefly, ChIP was carried out on $6x10^7$ cells using antibody against YY1 (Santa Cruz Biotechnology sc-1703). ChIP DNA was used to prepare libraries which were sequenced on the Illumina Genome Analyzer. Density profiles were generated as described (Mikkelsen, Ku et al. 2007). Promoters (RefSeq; http://genome.ucsc.edu) were classified as positive for YY1, H3K4me3 or H3K27me3 if the read density was significantly enriched ($p < 10^{-3}$) over a background distribution based on randomized reads generated separately for each dataset to account for the varying degrees of sequencing depth. ChIP-Seq data for YY1 are deposited to the NCBI GEO database under the following accession number GSE25197 (http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE25197).

Sites of Ezh2 enrichment ($p < 10^{-3}$) were calculated genomewide using sliding 1 kb windows, and enriched windows within 1 kb were merged.

YY1 motifs were identified using the MAST algorithm (Bailey and Gribskov 1998) where a match to the consensus motif was defined at significance level 5×10^{-5}. Motifs shown in Figure 2.4 are from UCSCs TFBS conserved track.

ACKNOWLEDGEMENTS

We thank A. Goren, A. Meissner, J. Rinn, T. Mikkelsen, M. Guttman, E. Lander and N. Shoresh for helpful discussions. We acknowledge L. Zagachin and the MGH RT-PCR core for technical assistance with quantitative PCR, as well as N. Geijsen for assistance with cell culture. We thank the staff of the Broad Institute Genome Sequencing Platform for assistance with reagents and data generation. We acknowledge David Conner of Harvard Medical School for recombineering and transgenic cell assistance.

REFERENCES

Atchison, L., A. Ghias, et al. (2003). "Transcription factor YY1 functions as a PcG protein in vivo." Embo J **22**(6): 1347-1358.

Azuara, V., P. Perry, et al. (2006). "Chromatin signatures of pluripotent cell lines." Nat Cell Biol **8**(5): 532-538.

Bailey, T. L. and M. Gribskov (1998). "Combining evidence using p-values: application to sequence homology searches." Bioinformatics **14**(1): 48-54.

Bernstein, B., T. Mikkelsen, et al. (2006). "A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells." Cell **125**: 315-326.

Boyer, L. A., K. Plath, et al. (2006). "Polycomb complexes repress developmental regulators in murine embryonic stem cells." Nature **441**(7091): 349-353.

Bracken, A. P., N. Dietrich, et al. (2006). "Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions." Genes Dev **20**(9): 1123-1136.

Cao, R., L. Wang, et al. (2002). "Role of histone H3 lysine 27 methylation in Polycomb-group silencing." Science **298**(5595): 1039-1043.

Creyghton, M. P., S. Markoulaki, et al. (2008). "H2AZ is enriched at polycomb complex target genes in ES cells and is necessary for lineage commitment." Cell **135**(4): 649-661.

Czermin, B., R. Melfi, et al. (2002). "Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites." Cell **111**(2): 185-196.

Dejardin, J., A. Rappailles, et al. (2005). "Recruitment of Drosophila Polycomb group proteins to chromatin by DSP1." Nature **434**(7032): 533-538.

Garcia, E., C. Marcos-Gutierrez, et al. (1999). "RYBP, a new repressor protein that interacts with components of the mammalian Polycomb complex, and with the transcription factor YY1." Embo J **18**(12): 3404-3418.

Goldberg, A. D., L. A. Banaszynski, et al. "Distinct factors control histone variant H3.3 localization at specific genomic regions." Cell **140**(5): 678-691.

Hansen, K. H., A. P. Bracken, et al. (2008). "A model for transmission of the H3K27me3 epigenetic mark." Nat Cell Biol **10**(11): 1291-1300.

Kanhere, A., K. Viiri, et al. "Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2." Mol Cell **38**(5): 675-688.

Kim, H., K. Kang, et al. (2009). "AEBP2 as a potential targeting protein for Polycomb Repression Complex PRC2." Nucleic Acids Res **37**(9): 2940-2950.

Kim, J. D., K. Kang, et al. (2009). "YY1's role in DNA methylation of Peg3 and Xist." Nucleic Acids Res **37**(17): 5656-5664.

Kim, T. G., J. C. Kraus, et al. (2003). "JUMONJI, a critical factor for cardiac development, functions as a transcriptional repressor." J Biol Chem **278**(43): 42247-42255.

Ko, C. Y., H. C. Hsu, et al. (2008). "Epigenetic silencing of CCAAT/enhancer-binding protein delta activity by YY1/polycomb group/DNA methyltransferase complex." J Biol Chem **283**(45): 30919-30932.

Ku, M., R. P. Koche, et al. (2008). "Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains." PLoS Genet **4**(10): e1000242.

Kuzmichev, A., K. Nishioka, et al. (2002). "Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein." Genes Dev **16**(22): 2893-2905.

Lee, T. I., R. G. Jenner, et al. (2006). "Control of developmental regulators by Polycomb in human embryonic stem cells." Cell **125**(2): 301-313.

Li, G., R. Margueron, et al. "Jarid2 and PRC2, partners in regulating gene expression." Genes Dev **24**(4): 368-380.

Liu, H., M. Schmidt-Supprian, et al. (2007). "Yin Yang 1 is a critical regulator of B-cell development." Genes Dev **21**(10): 1179-1189.

Lorente, M., C. Perez, et al. (2006). "Homeotic transformations of the axial skeleton of YY1 mutant mice and genetic interaction with the Polycomb group gene Ring1/Ring1A." Mech Dev **123**(4): 312-320.

Margueron, R., N. Justin, et al. (2009). "Role of the polycomb protein EED in the propagation of repressive histone marks." Nature **461**(7265): 762-767.

Mikkelsen, T. S., M. Ku, et al. (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." Nature **448**(7153): 553-560.

Mohn, F., M. Weber, et al. (2008). "Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors." Mol Cell **30**(6): 755-766.

Negre, N., J. Hennetin, et al. (2006). "Chromosomal distribution of PcG proteins during Drosophila development." PLoS Biol **4**(6): e170.

Pasini, D., P. A. Cloos, et al. "JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells." Nature **464**(7286): 306-310.

Peng, J. C., A. Valouev, et al. (2009). "Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells." Cell **139**(7): 1290-1302.

Ramirez-Carrozzi, V. R., D. Braas, et al. (2009). "A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling." Cell **138**(1): 114-128.

Ringrose, L. and R. Paro (2007). "Polycomb/Trithorax response elements and epigenetic memory of cell identity." Development **134**(2): 223-232.

Rinn, J. L., M. Kertesz, et al. (2007). "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs." Cell **129**(7): 1311-1323.

Schuettengruber, B., M. Ganapathi, et al. (2009). "Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos." PLoS Biol **7**(1): e13.

Schwartz, Y. B., T. G. Kahn, et al. (2006). "Genome-wide analysis of Polycomb targets in Drosophila melanogaster." Nat Genet **38**(6): 700-705.

Schwartz, Y. B. and V. Pirrotta (2007). "Polycomb silencing mechanisms and the management

of genomic programmes." Nat Rev Genet **8**(1): 9-22.

Shen, X., W. Kim, et al. (2009). "Jumonji modulates polycomb activity and self-renewal versus differentiation of stem cells." Cell **139**(7): 1303-1314.

Simon, J., A. Chiang, et al. (1993). "Elements of the Drosophila bithorax complex that mediate repression by Polycomb group products." Dev Biol **158**(1): 131-144.

Sing, A., D. Pannell, et al. (2009). "A vertebrate Polycomb response element governs segmentation of the posterior hindbrain." Cell **138**(5): 885-897.

Sui, G., B. Affar el, et al. (2004). "Yin Yang 1 is a negative regulator of p53." Cell **117**(7): 859-872.

Tanay, A., A. H. O'Donnell, et al. (2007). "Hyperconserved CpG domains underlie Polycomb-binding sites." Proc Natl Acad Sci U S A **104**(13): 5521-5526.

Tate, C. M., J. H. Lee, et al. (2009). "CXXC Finger Protein 1 Contains Redundant Functional Domains That Support Embryonic Stem Cell Cytosine Methylation, Histone Methylation, and Differentiation." Mol Cell Biol.

Tevosian, S. G., K. H. Albrecht, et al. (2002). "Gonadal differentiation, sex determination and normal Sry expression in mice require direct interaction between transcription partners GATA4 and FOG2." Development **129**(19): 4627-4634.

Tolhuis, B., E. de Wit, et al. (2006). "Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster." Nat Genet **38**(6): 694-699.

Tsai, M. C., O. Manor, et al. "Long noncoding RNA as modular scaffold of histone modification complexes." Science **329**(5992): 689-693.

Wang, L., J. L. Brown, et al. (2004). "Hierarchical recruitment of polycomb group silencing complexes." Mol Cell **14**(5): 637-646.

Woo, C. J., P. V. Kharchenko, et al. "A region of the human HOXD cluster that confers polycomb-group responsiveness." Cell **140**(1): 99-110.

Xi, H., Y. Yu, et al. (2007). "Analysis of overrepresented motifs in human core promoters

reveals dual regulatory roles of YY1." Genome Res **17**(6): 798-806.

Yue, R., J. Kang, et al. (2009). "Beta-arrestin1 regulates zebrafish hematopoiesis through binding to YY1 and relieving polycomb group repression." Cell **139**(3): 535-546.

Zhao, J., B. K. Sun, et al. (2008). "Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome." Science **322**(5902): 750-756.

Chapter 3:

DamID-seq, a New Method for

Global Characterization of Chromatin Regulators

DamID-seq, a New Method for Global Characterization of Chromatin Regulators

Vicky W. Zhou[1,2,3,4], Daniel Fernandez[3,5], Yoshiko Mito[1,2,3], Jun S. Liu[5], Bradley E. Bernstein[1,2,3]

1. Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA, 02114.

2. Center for Systems Biology and Center for Cancer Research, Massachusetts General Hospital, Boston, Massachusetts, USA, 02114.

3. Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 02142.

4. Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA, 02115.

5. Statistics Department, Harvard University, Cambridge, Massachusetts, USA, 02138.

AUTHOR CONTRIBUTIONS

V.W.Z. and B.E.B. conceived and designed the DamID-seq method, as presented in Figures 3.1 and 3.2. V.W.Z. performed all experiments, as shown in Figures 3.3A and 3.5. V.W.Z., D.M., and Y.M. designed the analyses. D.M. created the scatter plots in Figures 3.3B and 3.4.

ABSTRACT

Next-generation sequencing has greatly advanced our ability to study chromatin on a genomewide scale, and has been applied to globally characterize DNA methylation, histone modifications, and transcription factor binding, among other properties. However, we currently lack an ideal comprehensive assay for identifying the genomewide binding patterns of chromatin proteins, as current methods are hindered by difficulty crosslinking loosely or transiently bound proteins and poor antibodies. Here, we adapt a method for mapping chromatin regulators that uses a fusion enzyme and that does not rely on crosslinking and antibodies, and show that it can be used to globally map chromatin proteins in both human K562 and 293T cells.

INTRODUCTION

The human genome, at the most basic level, is a sequence of over three billion base pairs that is identical in virtually every cell of the human body. As the adult human body consists of over 100 trillion cells that correspond to at least 200 histologically unique cell types, the establishment and maintenance of these discrete and stable states is key to cell identity and differentiation. Cell state may be established and maintained by numerous factors, including DNA methylation, transcription factors, and chromatin proteins (Zhou, Goren et al. 2011). In recent years, technological advancements in next-generation sequencing have enabled many of these factors to be mapped genomewide.

While existing methods for genomewide mapping of factor binding work well for histone modifications and transcription factors, they have limited applicability to chromatin proteins.

Figure 3.1. Comparison Between ChIP-seq and DamID-seq Technologies

Schematic overview of steps for ChIP-seq and DamID-seq. (A) In ChIP-seq, a bound protein (dark blue) is crosslinked to DNA by formaldehyde. The crosslinked sample is sonicated until the DNA is 100-500bp in length. Then, the protein, along with the chromatin crosslinked to it, is pulled-down with an antibody that has been incubated with Protein A or G beads. The crosslinks are reversed by heat, and the DNA is purified. The DNA is sequenced, aligned to the genome, and viewed with the Integrative Genomics Viewer, which displays a histogram corresponding to the number of sequence reads at each genomic location. The shaded boxes indicate steps that often do not work well for chromatin proteins. (B) In DamID-seq, a bound protein (dark blue) is fused to DNA adenine methyltransferase (light blue). The Dam methylates adenines at nearby GATC sites (red). The methylated DNA is isolated, and the methylated regions are amplified by PCR. Note that after PCR, the DNA is no longer methylated, but in this schematic, we included the methylation marks for simplicity. The DNA fragments are then sonicated to 100-500bp in length, and the originally methylated ends are pulled-down with beads. The DNA is sequenced, aligned to the genome, and viewed with the Integrative Genomics Viewer, which displays a histogram corresponding to the number of sequence reads at each GATC site.

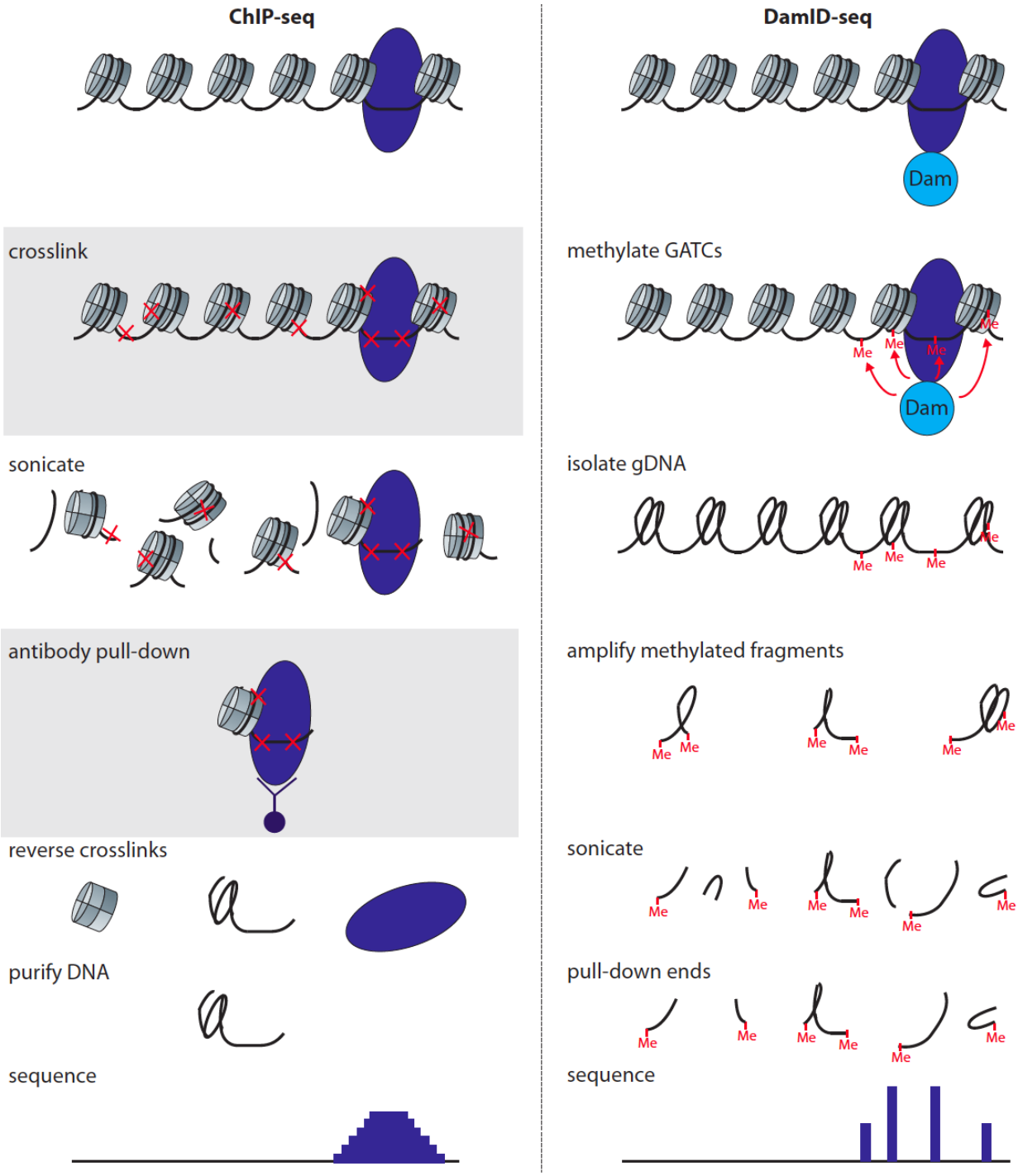

Figure 3.1 (Continued).

Specifically, chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is limited by poor crosslinking of loosely or transiently bound proteins, the lack of high-quality antibodies, poor solubility of compact chromatin, and the high sequencing depth necessary for broadly bound proteins (Figure 3.1). Although recent studies have begun to overcome these limitations through new chromatin immunoprecipitation procedures and antibody screening (Ram, Goren et al. 2011; Rhee and Pugh 2011), the vast majority of chromatin regulators remain uncharted. A complementary method that does not rely on antibodies and formaldehyde crosslinking, and that is based on a reduced representation of the genome, is needed to thoroughly address the binding of chromatin proteins.

In 2000, Bas van Steensel and Steven Henikoff presented a method called DamID for identifying binding sites of chromatin proteins by using a tethered *E.coli* DNA adenine methyltransferase (Dam) (van Steensel and Henikoff 2000). This method does not require crosslinking or antibodies, and samples the genome at GATC sites. They and others then used microarrays to create genomewide maps from these DamID libraries for various *Drosophila* proteins (van Steensel, Delrow et al. 2001; Greil, van der Kraan et al. 2003; Orian, van Steensel et al. 2003; Sun, Chen et al. 2003; van Steensel, Delrow et al. 2003; Bianchi-Frias, Orian et al. 2004; de Wit, Greil et al. 2005). In 2006, the van Steensel lab reported the application of DamID-microarray to mammalian cells (Vogel, Peric-Hupkes et al. 2007). Using this technology, they were able to map both CBX1 and nuclear lamina in mammalian cells (Vogel, Guelen et al. 2006; Guelen, Pagie et al. 2008).

Here, we adapted DamID for use with high-throughput sequencing, which we call DamID-seq (Figure 3.1). We used DamID-seq to profile the genomewide binding of chromatin proteins in both human K562 and 293T cells. We anticipate that this technology can complement

existing methods to generate a complete picture of chromatin in various cell types.

RESULTS

DamID is Adapted for High-Throughput Sequencing

To develop a new technology for genomewide mapping of mammalian chromatin proteins that addresses the limitations of ChIP-seq, we adapted DamID for high-throughput sequencing. DamID is based on fusing a protein of interest with *E. coli* DNA adenine methyltransferase (Dam), which methylates adenines at GATC sites (Vogel, Peric-Hupkes et al. 2007). This fusion protein is inserted into mammalian cells by lentiviral infection. Inside the cell, the fusion protein binds DNA and marks nearby GATC sequences with adenine methylation (Figure 3.1). By isolating the genomic DNA (gDNA) and amplifying the methylated regions, one can map the binding sites for the protein of interest.

To date, all published mammalian DamID libraries have been queried by microarray. For instance, DamID followed by microarray has been used to map CBX1 in MCF7 human breast carcinoma cell lines (Vogel, Guelen et al. 2006), laminB1 in human fibroblasts (Guelen, Pagie et al. 2008), and 53 chromatin proteins in *Drosophila* (Filion, van Bemmel et al. 2010).

There are several advantages to sequencing DamID libraries over using microarrays: sequencing offers quantitative digital data and no hybridization bias. To capitalize on the benefits of sequencing, we designed a method for amplifying methylated fragments that enables capture of each individual methylated GATC site. Sequencing such a library reveals counts for methylation of every GATC site, from which the binding pattern of the protein can be inferred.

Figure 3.2. DamID-seq Technology

Schematic of steps for adapting DamID for high-throughput sequencing. Methylated DNA is digested with *DpnI*, ligated with adaptors, amplified by PCR with a biotinylated primer (blue circle), fragmented with Covaris, pulled-down by streptavidin beads (yellow circle), digested with *DpnII*, and prepared for high-throughput sequencing.

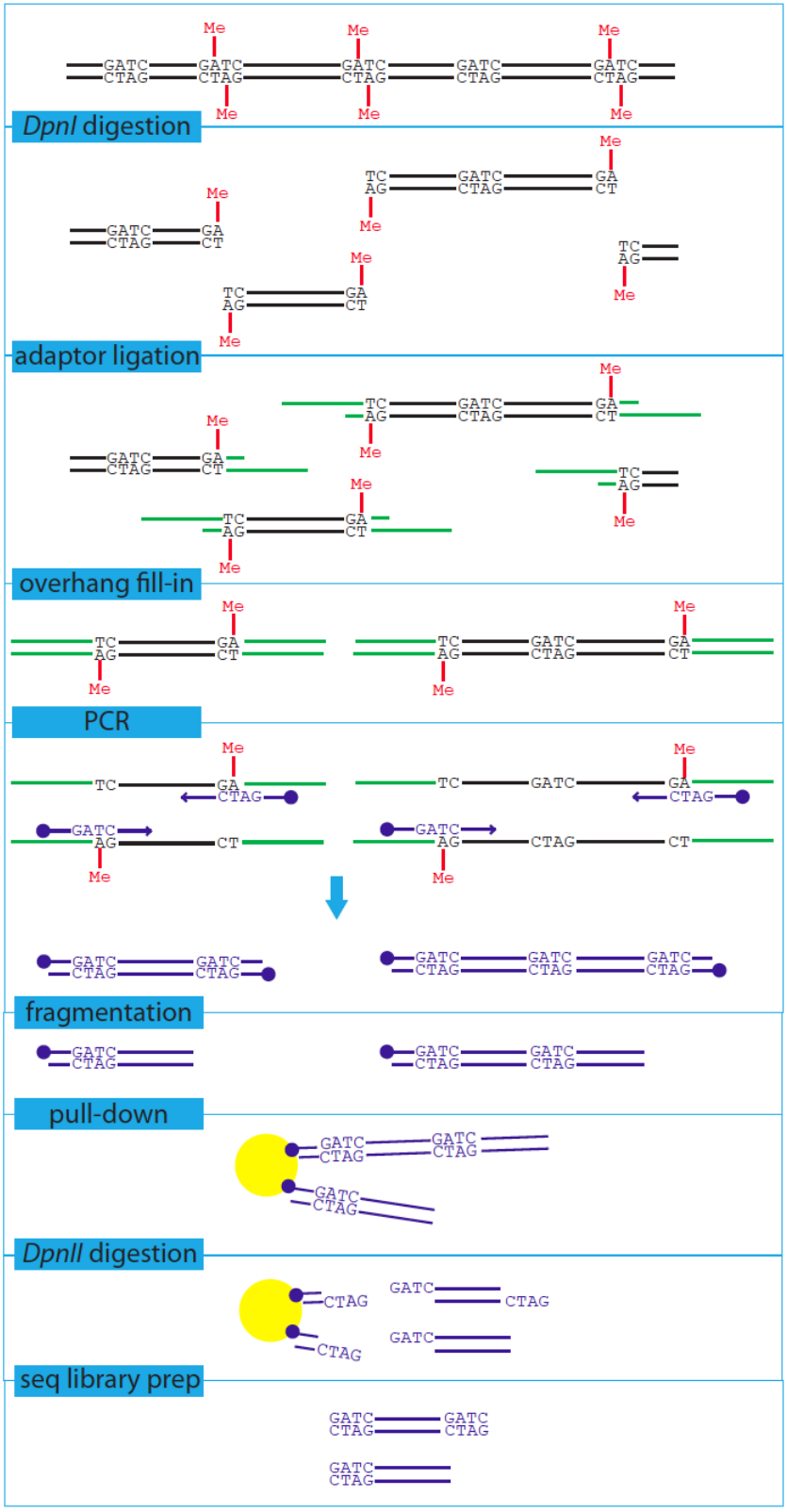

Figure 3.2 (Continued).

Specifically, we amplified methylated fragments by first digesting gDNA with *DpnI*, which cuts methylated GATC sites, and ligating adaptors at these cut ends (Figure 3.2). These fragments are then amplified with a biotinylated primer, yielding fragments ranging from 0.4-3kb in size. To shorten these fragments to a size amenable to sequencing, they are sonicated using Covaris until 100-500bp in size.

Since we only want to sequence methylated GATC sites, which are located at the biotinylated ends of these fragments, we pulled down such fragments with streptavidin beads. The DNA was cut off from the beads with *DpnII*, which digests GATC sites regardless of methylation status. The resulting sequencing library consists of GATC sites followed by the flanking DNA sequence that can be used to map these sites to the human genome.

Validation of DamID-seq with Replicates

We used DamID-seq to map the genomewide binding of a chromatin protein (CBX8) and a nuclear lamina protein (LMNB1) in K562 cells (Figure 3.3A). We generated over 22 million reads for CBX8, and over 20 million reads for LMNB1. We filtered for reads that start with GATC, which resulted in over 9 million and over 8 million reads for CBX8 and LMNB1, respectively. If our pull-down was 100% efficient, we would expect 50% of the reads to start with GATCs, so our pull-downs appear to be 80-90% efficient. Our GATC-starting read numbers corresponds to 1.3x and 1.1x coverage of GATC sites in the genome. While this coverage is sufficient because our enrichment method isolates but a small fraction of such sites, we note that additional sequencing reads, and thereby greater coverage, can improve the sensitivity of our method.

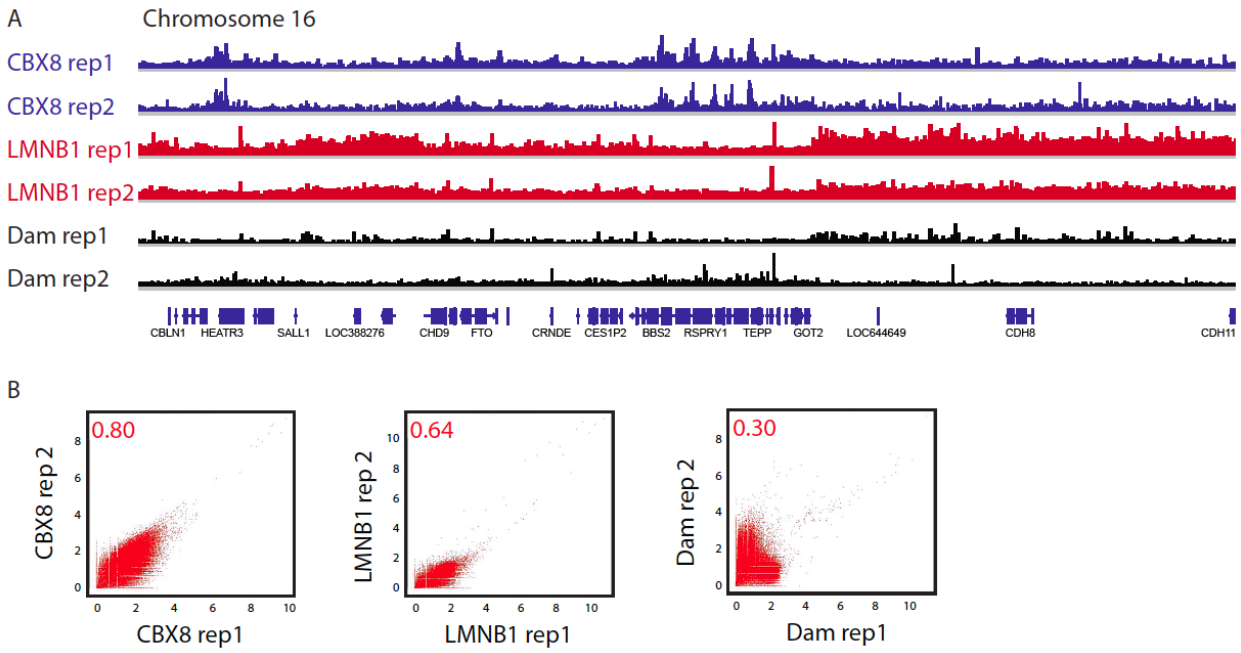

Figure 3.3. Validation of DamID-seq with Replicates

(A) DamID-seq tracks of two replicates each of CBX8 (blue), LMNB1 (red), and Dam only (black) in K562 cells.

(B) Scatter plots comparing 2 kb bins of reads between the two replicates of CBX8, the two replicates of LMNB1, and the two replicates of Dam only. The red number in the upper left corner is the correlation.

We performed two biological replicates each for CBX8 and LMNB1 (Figure 3.3A). The resulting maps clearly show consistent binding peaks for both proteins. CBX8 peaks appear punctate, as would be expected by its role as a member of the Polycomb repressive complex 1 (Beisel and Paro 2011). LMNB1, on the other hand, binds to broad Megabase domains, which is consistent with previous reports (Guelen, Pagie et al. 2008).

We then quantitatively compared the two biological replicates with each other by comparing read numbers found in 2kb windows across the genome (Figure 3.3B). We found that the CBX8 replicates had a correlation coefficient of 0.8, and that the LMNB1 replicates had a correlation coefficient of 0.64. These are reasonably high values for distinct biological samples, and validates that DamID-seq is a robust method. CBX8 replicates likely have a higher correlation coefficient than LMNB1 replicates because it has more total sequencing reads, and therefore greater coverage. Furthermore, since CBX8 peaks are narrower than that of LMNB1, its effective coverage is even greater.

Finally, we conducted a control to test for potential false positives resulting from the background binding of the Dam protein or PCR duplicates. We mapped the genomewide binding of the Dam protein by itself in two biological replicates (Figure 3.3A). Both replicates show that the Dam protein binds at a basal level across the region, with a few modest peaks that correspond to either regions that Dam preferentially binds or PCR duplicates. However, these patterns do not account for the strong binding peaks found in the CBX8 or LMNB1 maps, indicating that those peaks are primarily due to the binding of the protein of interest.

We also quantitatively compared the two Dam replicates with each other and found a correlation coefficient of 0.30 (Figure 3.3B). This low correlation is expected for a control protein that binds at a low level across the entire genome, and validates that the Dam protein by

itself does not have notable binding preferences.

Limited Bias in DamID-seq

We sought to identify and quantify the sources of bias in DamID-seq data. We first examined bias resulting from the GATC distribution, since reads must map to GATC sites. For 2kb bins across one representative chromosome, we compared the number of GATCs with the number of DamID-seq reads by plotting a log-scale density scatter plot (Figure 3.4A). We examined both reads from the Dam only map, and reads from the CBX8 map. We found that there were strong correlations of 0.70 and 0.61, respectively, between the number of GATC sites and the number of reads, which is expected given our technology.

Next, we asked whether we could counter this GATC bias by normalizing the reads by the number of GATC sites per 2kb bin and again plotting a log-scale density scatter plot (Figure 3.4B). Indeed, there is a low correlation between the number of GATC sites and the normalized number of reads. The correlation coefficient for Dam only data is 0.29, and for CBX8 data, 0.23. The limited remaining GATC bias may result from the fact that two methylated GATC sites are required to amplify a given region. Since our PCR fragments appear to range from 400bp - 3kb, we may miss potential positive regions if there are fewer than two GATC sites within 3kb.

We then examined bias resulting from the G+C content of the DNA sequence, which can affect PCR amplification and sequencing. For 2kb bins across one representative chromosome, we compared the G+C percentage with the log of the number of DamID-seq reads by plotting a density scatter plot (Figure 3.4C-D).

Figure 3.4. Limited Bias in DamID-seq

(A) For 2 kb bins across chromosome 19, density scatter plot comparing the natural log of number of GATCs with the natural log of Dam reads (left) or CBX8 reads (right). A low density of points is colored blue, and a high density of points is colored red. The correlation coefficient is indicated in red at the upper left-hand corner. (B) For 2 kb bins across chromosome 19, density scatter plot comparing the natural log of number of GATCs with the natural log of Dam reads normalized by the number of GATCs (left) or CBX8 reads normalized by the number of GATCs (right). A low density of points is colored blue, and a high density of points is colored red. The correlation coefficient is indicated in red at the upper right-hand corner. (C) For 2 kb bins across chromosome 19, density scatter plot comparing the G+C percentage with the natural log of Dam reads (left) or CBX8 reads (right), for G+C percentages above the median. A low density of points is colored blue, and a high density of points is colored red. The correlation coefficient is indicated in red at the upper right-hand corner. (D) For 2 kb bins across chromosome 19, density scatter plot comparing the G+C percentage with the natural log of Dam reads (left) or CBX8 reads (right), for G+C percentages below the median. A low density of points is colored blue, and a high density of points is colored red. The correlation coefficient is indicated in red at the upper left-hand corner.

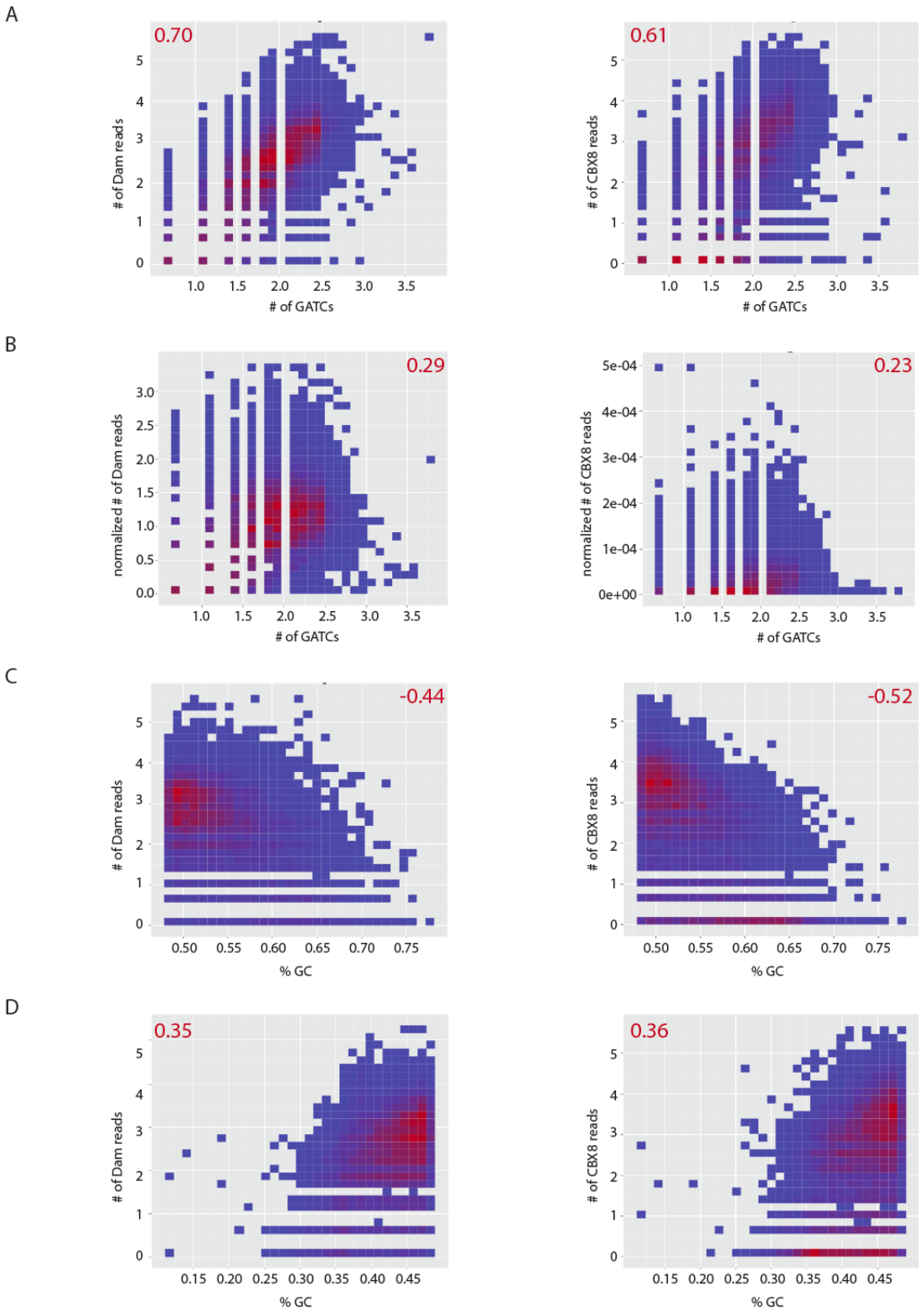

Figure 3.4 (Continued).

We found that the G+C bias was nonlinear, so we split the plots into bins with G+C content above the median (Figure 3.4C) and bins with G+C content below the median (Figure 3.4D). For high G+C bins, there is a negative correlation between G+C percentage and number of reads: namely, -0.44 for Dam only data and -0.52 for CBX8 data. In contrast, for low G+C bins, there is a modest positive correlation between G+C percentage and number of reads: namely, 0.35 for Dam only data and 0.36 for CBX8 data. This is comparable to the G+C bias found in ChIP-seq data (Goren, Ozsolak et al. 2010). Thus, DamID-seq appears to have a limited GATC bias and G+C bias that is comparable to existing technology.

DamID-seq in Multiple Human Cell Types

We have used DamID-seq to map chromatin proteins in two human cell lines: K562 cells, which are suspension cells derived from myelogenous leukemia (Figure 3.3), and 293T cells, which are adherent cells derived from embryonic kidney (Figure 3.5). To apply DamID-seq to different cell types, we optimized the lentiviral infection protocol for each case. We found that the remainder of the protocol works well regardless of the cell type. Thus, we predict that DamID-seq can be used to map as many human cell types as can be infected by lentivirus.

For the map of chromatin protein CBX1 in 293T cells, we conducted one additional control to test for potential false negatives resulting from the PCR step. Specifically, the PCR step amplifies regions between methylated GATC sites, even if there is an unmethylated GATC site in between (Figure 3.2). We reasoned that the chance of this occurring is modest because the Dam protein has a reach of 1kb from its binding site, while the PCR products are only up to 3kb in length (van Steensel and Henikoff 2000).

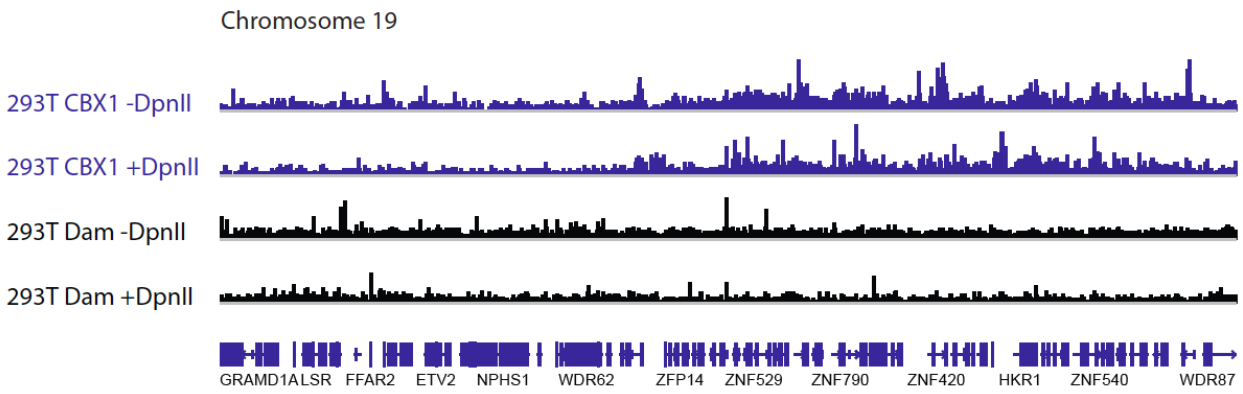

Figure 3.5. DamID-seq in 293T Cells

DamID-seq tracks of CBX1 and Dam only in 293T cells. The "-DpnII" maps were constructed according to the general protocol, as outlined in Figure 3.2. The "+DpnII" maps were constructed with an additional *DpnII* digestion after the adaptor ligation, and before the overhang fill-in and PCR amplification.

We experimentally verified this by introducing a *DpnII* digestion after the adaptor ligation, and before the overhang fill-in and PCR amplification. We found that with or without this additional *DpnII* digestion, the resulting map reveals similar broad binding peaks (Figure 3.5). Thus, we concluded that there were no major false positives resulting from the PCR step.

DISCUSSION

Here, we adapted DamID for high-throughput sequencing to create genomewide maps of chromatin proteins. This method circumvents the need for crosslinking and antibodies, and is particularly suited for broadly, loosely, and transiently bound proteins. We envision that DamID-seq will complement ChIP-seq, which is better suited for mapping histone modifications, transcription factors and other proteins that intimately interact with the DNA. This approach should help to build a complete picture of chromatin in mammalian cells.

We note that for proteins that can be readily mapped by both ChIP-seq and DamID-seq, ChIP-seq may be the better choice, as it profiles the endogenous protein over 10 minutes at a 25bp resolution. DamID-seq, on the other hand, profiles a lowly expressed exogenous fusion protein over 3 days at a 1kb resolution. However, for proteins that cannot be mapped by ChIP-seq, such as the nuclear lamina and other broad proteins in regions of compact chromatin or without high-quality antibodies, DamID-seq may be the only option for obtaining a genomewide map of binding.

We further note that DamID-seq is uniquely capable of mapping mutant proteins, which will be essential for studying disease states. During the cloning step of the DamID-seq protocol, the sequence of the protein of interest can be mutated, so that when it is expressed in the cell, a

mutant protein will be fused to the Dam protein. We can then generate a map of the mutant protein's binding, and compare it to a map of the wild-type protein's binding to identify aberrant patterns. This could further our understanding of the many diseases that involve mutant DNA-binding proteins.

To use DamID-seq to globally characterize mammalian chromatin, we need to develop a peak caller to interpret binding regions from the maps, and to map additional chromatin proteins to begin to illuminate genomewide patterns. We will address these issues in the next Chapter.

METHODS

Plasmid Construction

We obtained plasmids pLgw V5-EcoDam (Dam only negative control), pLgw EcoDam-V5-RFC1 (N-terminus Dam vector), pLgw RFC1-V5-EcoDam (C-terminus Dam vector), and pLgw CBX1-V5-EcoDam (CBX1-Dam positive control) from the Bas van Steensel laboratory. We obtained ORFs in plasmid pDONR221 (or pDONR201) for CBX8 (cloneID HsCD00079972) and LMNB1 (clone ID HsCD00043675) from the PlasmID collection at the Dana-Farber/Harvard Cancer Center DNA Resource Core.

We used Invitrogen Gateway® Cloning technology to clone these ORFs into either pLgw EcoDam-V5-RFC1 (N-terminus Dam vector) or pLgw RFC1-V5-EcoDam (C-terminus Dam vector), depending on whether the available ORF had a stop codon. Namely, LMNB1 was cloned with an N-terminus Dam, and CBX8 was cloned with a C-terminus Dam.

Cell Culture

293T cells were grown according to standard protocols in Gibco KO DMEM media supplemented with 10% fetal bovine serum (FBS, Atlas Biologicals, F-0500-A), 1% Penicillin/Streptomycin (Invitrogen, 15140122), and 1% Glutamax. K562 erythrocytic leukemia cells (ATCC CCL-243) were grown according to standard protocols in RPMI 1640 media (Invitrogen, 22400105) supplemented with 10% fetal bovine serum (FBS, Atlas Biologicals, F-0500-A) and 1% Penicillin/Streptomycin (Invitrogen, 15140122).

Lentiviral Production

293T cells were grown in 15cm dishes until 60-80% confluence. 1140uL of DMEM was combined with 60uL of Fugene. After 5 min, the following three plasmids were added: Gag, pol and rev plasmid (6ug), VSV envelope plasmid (3ug), and specific cloned Dam-protein plasmid or GFP (9ug). This mixture was incubated at room temperature for 5 min, then added dropwise to the 293T cells. After 8-12 hours, the media was replaced with 12mL fresh medium. After 72 hours, the virus was collected filtered through 0.45 uM. For K562 infections, the virus was ultracentrifuged at 28000 rpm for 2 hours in an SW41Ti rotor at 4°C. The virus was resuspended in 100uL PBS, and left at 4°C overnight.

293T Lentiviral Infection

293T cells were plated in two wells of a 6-well plate and grown until 70% confluency for

each experiment. For each well, 1.5uL Polybrene (10mg/mL) was added and incubated for 30min, before adding 0.75mL of unconcentrated virus and 0.75mL of media. The cells were then returned to 37°C overnight.

The next day, 2mL fresh media was added per well to dilute out the virus. After 48 hours later, the cells were harvested and the gDNA was isolated using the Qiagen DNA Micro Kit, "Isolation of gDNA from Small Volumes of Blood" protocol. The DNA was eluted in 200uL buffer AE, and quantified by Nanodrop.

K562 Lentiviral Infection

K562 cells were counted with a hemacytometer and 1.5 million cells were allocated per each infection. Each aliquot of cells was spun down and resuspended in 3mL fresh media, and plated in one well of a 6-well plate. 2uL Polybrene (10mg/mL) and 30uL of concentrated virus were added. The cells were spin-infected at 2500 rpm, for 90 min, at room temperature. The cells were then returned to 37°C overnight.

The following day, the infected cells were spun down and resuspended in 3mL fresh media. After 48 hours later, the cells were harvested and the gDNA was isolated using the Qiagen DNA Micro Kit, "Isolation of gDNA from Small Volumes of Blood" protocol. The DNA was eluted in 200uL buffer AE, and quantified by Nanodrop.

DamID Library Preparation and Sequencing

The gDNA was ethanol precipitated, and dissolved in TE pH7.5 to a concentration of 1

ug/uL. 2.5uL of gDNA was digested with 0.5uL of DpnI (NEB, 20 U/uL) at 37°C overnight in PCR tubes. DpnI was inactivated by heating to 80°C for 20 min. The DpnI-digested gDNA was ligated with the adaptor AdR, which is made by mixing and slowly annealing AdR-top (5' CTAATACGACTCACTATAGGGCAGCGTGGTCGCGGCCGAGGA 3') and AdR-bottom (5' TCCTCGGCCG 3'). This ligation was completed using 1uL T4 Ligase (Roche, 5U/uL) for 2 hours at 16°C. The T4 ligase was inactivated by heating to 65°C for 10 min. The resulting 20uL volume reaction was diluted to 50uL with ddH2O.

To amplify the regions flanked by adaptors, the following PCR was setup: 10uL DNA, 5uL 10x cDNA PCR reaction buffer (Clontech), 0.625uL primer bio-Adr-PCR (5' bio-GGTCGCGGCCGAGGATC 3', 100uM), 1uL dNTPs (10mM), 1uL PCR advantage enzyme mix (Clontech, 50X), 32.375uL ddH2O. The PCR reaction program was as follows: 1 cycle of 68°C (10min), 94°C (1min), 65°C (5min), 68°C (15min); 3 cycles of 94°C (1min), 65°C (1min), 68°C (10min); 17 cycles of 94°C (1min), 65°C (1min), 68°C (2min). 5uL of the PCR products were ran on a gel to verify successful digestion and amplification.

The PCR products were cleaned with the Qiagen MinElute PCR Purification kit, and eluted in 20uL ddH2O. Following quantification by Nanodrop, 3ug of each sample was diluted in 100uL ddH2O. These samples were sonicated with Covaris using the following settings: 10% duty cycle, 5 intensity, 200 cycles per burst, for 4.5 min total. 10uL of the Covaris-sonicated samples were ran on a gel to verify sonication to 100-500bp.

To pull down the biotinylated ends of the PCR products, we used Invitrogen Dynabeads® MyOne Streptavidin T1 beads. 50uL of these beads were washed three times with 50uL of 1X Binding and Washing (B&W) buffer (5mM Tris-HCl pH 7.5, 0.5mM EDTA, 1M NaCl). The washed beads were resuspended in 75uL of Covaris-sonicated DNA, 100uL of 2X B&W buffer,

and 25uL H2O. This mixture was incubated at 4°C for 15min on a rotator. Following a quick spin and decanting the supernatant, the beads were washed three times with 200uL of 1X B&W buffer.

The bound DNA was removed off the beads by digestion with DpnII at 37°C for 1 hour. The supernatant was collected and cleaned using the Qiagen MinElute Reaction Cleanup kit. The resulting DNA was eluted in 20uL H2O and quantified with Qubit. qPCR analysis was performed to validate the DamID DNA before submission for sequencing.

Libraries of DamID samples were prepared according to the Illumina Genomic DNA protocol, as described previously (Mikkelsen et al., 2007). The DamID-seq libraries were sequenced on Illumina GAII sequencers according to standard Illumina protocols.

DamID-seq Data Analysis

Sequence reads were aligned to the human genome reference (hg19). We filtered out low-quality reads, and filtered in reads that map to GATC sites. The number of reads was counted at each GATC site. The reads were viewed using the Integrative Genomics Viewer (http://www.broadinstitute.org/igv/).

ACKNOWLEDGMENTS

We thank B. van Steensel for the DamID vectors; M. Suva for the lentivirus vectors and protocol; O. Ram for the K562 cells; and T. Durham and N. Shoresh for help processing the raw sequence reads. V.W.Z. was supported by an NSF Graduate Research Fellowship and National

Defense Science and Engineering Graduate Fellowship. D.F. is advised and funded by the Jun S. Liu Laboratory. B.E.B. is a Charles E. Culpeper Medical Scholar and Early Career Scientist of the Howard Hughes Medical Institute. Research in the Bernstein Laboratory is supported by funds from the Burroughs Wellcome Fund, HHMI, and the NIH.

REFERENCES

Beisel, C. and R. Paro (2011). "Silencing chromatin: comparing modes and mechanisms." Nat Rev Genet **12**(2): 123-135.

Bianchi-Frias, D., A. Orian, et al. (2004). "Hairy transcriptional repression targets and cofactor recruitment in Drosophila." PLoS Biol **2**(7): E178.

de Wit, E., F. Greil, et al. (2005). "Genome-wide HP1 binding in Drosophila: developmental plasticity and genomic targeting signals." Genome Res **15**(9): 1265-1273.

Filion, G. J., J. G. van Bemmel, et al. (2010). "Systematic protein location mapping reveals five principal chromatin types in Drosophila cells." Cell **143**(2): 212-224.

Goren, A., F. Ozsolak, et al. (2010). "Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA." Nat Methods **7**(1): 47-49.

Greil, F., I. van der Kraan, et al. (2003). "Distinct HP1 and Su(var)3-9 complexes bind to sets of developmentally coexpressed genes depending on chromosomal location." Genes Dev **17**(22): 2825-2838.

Guelen, L., L. Pagie, et al. (2008). "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions." Nature **453**(7197): 948-951.

Orian, A., B. van Steensel, et al. (2003). "Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network." Genes Dev **17**(9): 1101-1114.

Ram, O., A. Goren, et al. (2011). "Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells." Cell **147**(7): 1628-1639.

Rhee, H. S. and B. F. Pugh (2011). "Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution." Cell **147**(6): 1408-1419.

Sun, L. V., L. Chen, et al. (2003). "Protein-DNA interaction mapping using genomic tiling path microarrays in Drosophila." Proc Natl Acad Sci U S A **100**(16): 9428-9433.

van Steensel, B., J. Delrow, et al. (2003). "Genomewide analysis of Drosophila GAGA factor target genes reveals context-dependent DNA binding." Proc Natl Acad Sci U S A **100**(5): 2580-2585.

van Steensel, B., J. Delrow, et al. (2001). "Chromatin profiling using targeted DNA adenine methyltransferase." Nat Genet **27**(3): 304-308.

van Steensel, B. and S. Henikoff (2000). "Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase." Nat Biotechnol **18**(4): 424-428.

Vogel, M. J., L. Guelen, et al. (2006). "Human heterochromatin proteins form large domains containing KRAB-ZNF genes." Genome Res **16**(12): 1493-1504.

Vogel, M. J., D. Peric-Hupkes, et al. (2007). "Detection of in vivo protein-DNA interactions using DamID in mammalian cells." Nat Protoc **2**(6): 1467-1478.

Zhou, V. W., A. Goren, et al. (2011). "Charting histone modifications and the functional organization of mammalian genomes." Nat Rev Genet **12**(1): 7-18.

Chapter 4:

DamID-seq Maps Genomewide Distribution of Polycomb and Heterochromatin Proteins

DamID-seq Maps Genomewide Distribution of Polycomb and Heterochromatin Proteins

Vicky W. Zhou[1,2,3,4], Daniel Fernandez[3,5], Yoshiko Mito[1,2,3], Oren Ram[1,2,3], Tim Durham[3], Noam Shoresh[3], Jun S. Liu[5], Bradley E. Bernstein[1,2,3]

1. Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA, 02114.

2. Center for Systems Biology and Center for Cancer Research, Massachusetts General Hospital, Boston, Massachusetts, USA, 02114.

3. Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 02142.

4. Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA, 02115.

5. Statistics Department, Harvard University, Cambridge, Massachusetts, USA, 02138.

AUTHOR CONTRIBUTIONS

V.W.Z. and B.E.B. conceived and designed the DamID-seq method. V.W.Z. performed all DamID-seq experiments, as shown in Figures 4.1E, 4.2A, and 4.3A. O.R. performed ChIP-seq experiments shown in Figure 4.2A.

V.W.Z., D.M., and Y.M. designed the analyses. V.W.Z. executed the analyses presented in Figures 4.1A, 4.2C, 4.2D, 4.5, and Table 4.1. V.W.Z. and D.F. executed the analyses presented in Figures 4.1B and 4.1D. D.F. coded and executed the analyses presented in Figures 4.1E, 4.1F, 4.1G, 4.2B, 4.3B, 4.4B, and 4.4C. T.D. and N.S. designed and executed the interim peak caller in Figure 4.1C. O.R. created the correlation matrix in Figure 4.4A.

ABSTRACT

Chromatin is a multi-layered structure composed of DNA, histones, and associated proteins. While histone modifications and transcription factors are readily mapped by ChIP-seq, DamID-seq can map chromatin proteins, which are often broadly, loosely and transiently bound. Here, we developed a DamID-seq peak caller, which identifies binding sites for chromatin proteins using a ratio between the protein map and a Dam only control map. We used DamID-seq and the peak caller to map the binding of 12 chromodomain-containing and related proteins in human K562 cells. We found that our proteins cluster into two modules: 1) Polycomb-related and 2) heterochromatin-related. Polycomb proteins bind developmental genes, while heterochromatin proteins CBX1, 3, and 5 bind broad olfactory receptor (OR) and zinc finger (ZNF) domains. Surprisingly, unlike other Polycomb proteins, CBX2 uniquely binds to genes involved with modifying proteins. Our findings advance the model that the genome is compartmentalized into domains, and identify the distinct protein components that associate respectively with Polycomb and heterochromatin domains in human cells.

INTRODUCTION

The mammalian genome is bound by hundreds of chromatin proteins that serve the dual purpose of structurally compacting the genome within the nucleus and functionally regulating its underlying DNA sequence. Many of these proteins work in combination and affect broad genomic regions, advancing a model in which the genome may be compartmentalized into higher-order domains.

Recent work lends support to a domain organization of the genome. Megabase (Mb) domains of H3K9me2 and lamina association sequester silenced heterochromatic regions (Guelen, Pagie et al. 2008; Wen, Wu et al. 2009). H3K27me3 blocks and Polycomb bodies may repress large sets of genes upon differentiation (Sexton, Schober et al. 2007; Pauler, Sloane et al. 2009). Furthermore, several large-scale profiling studies of chromatin proteins consistently reveal a limited number of domains: five chromatin "colors" in fly (Filion, van Bemmel et al. 2010), and six "modules" in human (Ram, Goren et al. 2011).

Despite the suggestion of the above studies that chromatin proteins can be divided into a relatively limited number of domains, biochemical and imaging studies on individual chromatin proteins have long emphasized their unique binding patterns. An exemplar is the eight human CBX proteins, which all contain one chromodomain that recognizes methylated histone lysines (Yap and Zhou 2011). Five of these proteins (CBX2, 4, 6, 7, 8) are homologous to *Drosophila* Pc (*d*Pc), and the other three (CBX1, 3, 5) are homologous to *Drosophila* HP1 (*d*HP1). Despite their similarities in structure, each CBX protein has a distinct *in vitro* peptide binding specificity (Kaustov, Ouyang et al. 2011), subnuclear distribution (Vincenz and Kerppola 2008), and binding partners (Lomberk, Wallrath et al. 2006; Rosnoblet, Vandamme et al. 2011; Vandamme, Volkel et al. 2011). However, a high-resolution, genomewide comparison of their *in vivo* binding patterns is yet to be elucidated.

Existing methods for mapping *in vivo* binding genomewide have limited applicability to chromatin proteins. In the previous Chapter, we adapted DamID for high-throughput sequencing to create DamID-seq, which relies on a fusion enzyme, rather than antibodies and formaldehyde crosslinking, and can address the binding of chromatin proteins, which are often broadly, loosely, or transiently bound.

Here, we used DamID-seq to map the genomewide binding of 12 chromatin proteins in human K562 cells. Specifically, we chose to leverage the power of DamID-seq by mapping proteins that are thought to be broadly and loosely bound to chromatin: CBX proteins, Polycomb repressive complex 1 (PRC1) and 2 (PRC2) components, and nuclear lamina. We also developed a new peak caller that enabled us to generate novel insights about the global binding patterns of these proteins in mammals.

Globally, we found that our panel of chromatin proteins clusters into two major modules: 1) Polycomb-related, and 2) heterochromatin-related. While Polycomb proteins bind to developmental genes, heterochromatin proteins CBX1, 3, and 5 bind to broad olfactory receptor (OR) and zinc finger (ZNF) protein clusters. CBX2 surprisingly exhibits unique binding patterns, different from either protein group. Our findings suggest that Polycomb and heterochromatin domains exhibit distinct properties, and further advance a domain model of the genome.

RESULTS

A New Peak Caller for DamID-seq Data

Since DamID-seq is a novel source of sequencing data, we could not use existing ChIP-seq peak callers to identify binding sites. One of the fundamental differences of DamID-seq is that only GATC sites in the genome carry measurable data. The distribution of GATC sites in the human genome is, in fact, not arbitrary. GATC sites are often less than 200bp apart (Figure 4.1A), while a randomly distributed four-base motif would be expected every 256bp.

Figure 4.1. A New Peak Caller for DamID-seq Data

(A) Distribution of distances between GATC sites in the human genome. The graph is truncated at 2000bp to show the distances with the highest frequencies. (B) Fold enrichment of GATC sites over the whole human genome for six aggregated state annotations. State annotations are based on nine histone modification maps in human K562 cells (Ernst, Kheradpour et al. 2011). (C) An interim peak caller that adapts Scripture for DamID by constructing a "GATC genome." The reads at GATC sites are concatenated into "chromosomes," demarcated by regions >5kb that lack GATC sites. Scripture is run on this "GATC genome," and enriched peaks are mapped back to the real genome. (D) Fold enrichment of Dam peaks, as called by the interim peak caller, over the whole human genome for six aggregated state annotations. State annotations are based on nine histone modification maps in human K562 cells (Ernst, Kheradpour et al. 2011). (E) A new log2 peak caller for DamID-seq data. DamID-seq tracks for RNF2, Dam control, and the log2 ratio of RNF2/Dam are shown. The peak caller is based on the log2 ratio, and the resulting peaks are below the tracks. (F) For 2 kb bins across chromosome 19, density scatter plot comparing the natural log of number of GATCs with the log2 peak caller reads for CBX8. A low density of points is colored blue, and a high density of points is colored red. The correlation coefficient is indicated in red at the upper left-hand corner. (G) For 2 kb bins across chromosome 19, density scatter plot comparing the G+C percentage with the log2 peak caller reads for CBX8. A low density of points is colored blue, and a high density of points is colored red. The correlation coefficient is indicated in red at the upper right-hand corner.

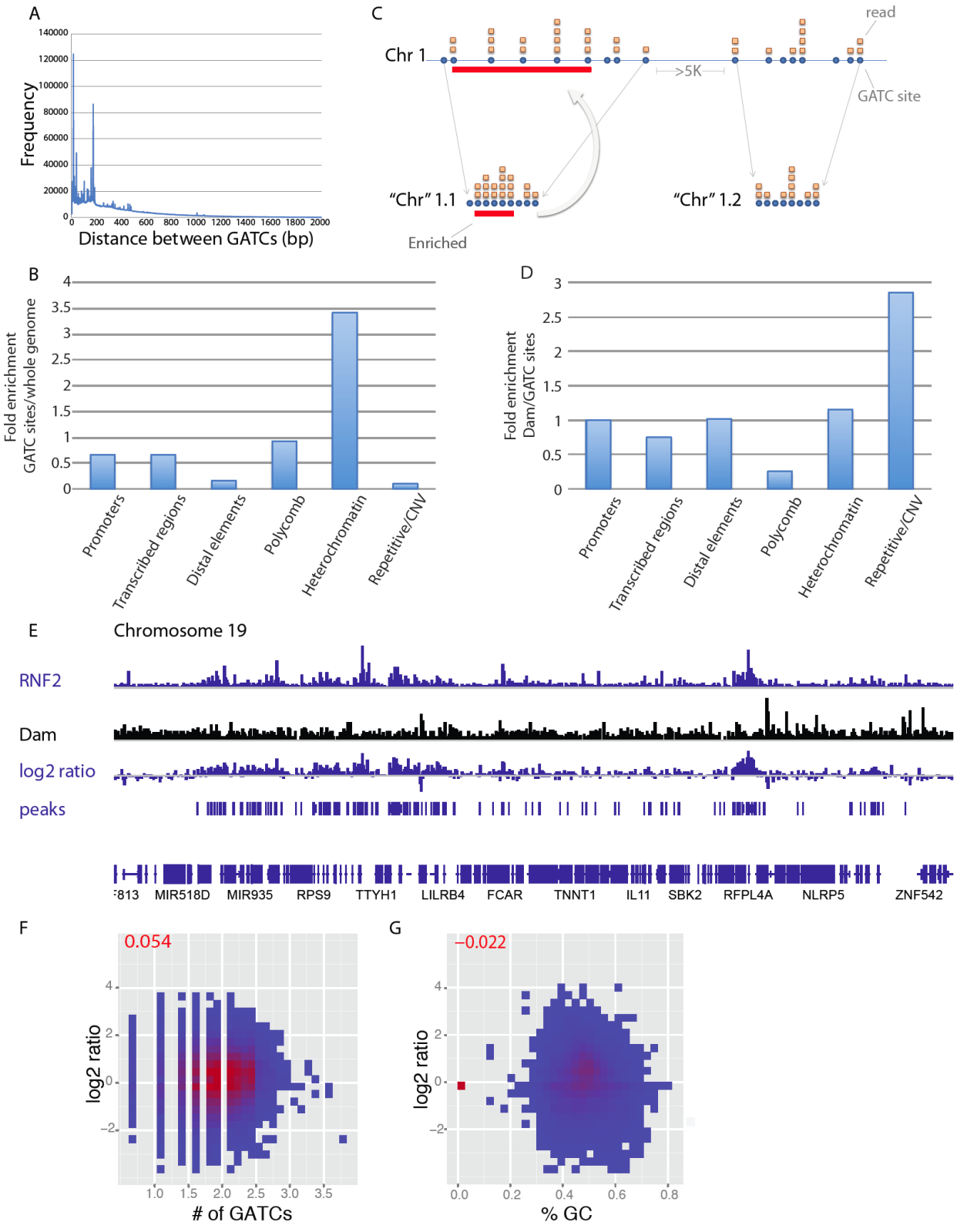

Figure 4.1 (Continued).

To examine the distribution of GATC sites at different regions of the genome, we used genomewide chromatin state annotations for human K562 cells, which was derived from nine histone modification maps (Ernst, Kheradpour et al. 2011). We compared the location of GATC sites with these state annotations, which we aggregated into "promoters," "transcribed regions," "distal elements," "Polycomb," "heterochromatin," and "repetitive/copy-number variants." We found that GATC sites are nearly 3.5-fold enriched at heterochromatin, while they are depleted from elsewhere in the genome (Figure 4.1B). Thus, DamID-seq effectively gives more representation at heterochromatin regions.

The second fundamental difference of DamID-seq is that the Dam protein itself may have an inherent bias for particular genomic regions (Vogel, Peric-Hupkes et al. 2007). To analyze the nature and extent of this Dam bias, we mapped the binding of the Dam protein alone using DamID-seq. We needed an interim peak caller to identify any enriched sites in this sample.

We adapted Scripture, a peak caller that has successfully identified ChIP-seq enriched regions (Guttman, Garber et al. 2010), for DamID-seq data by collapsing the genome to a series of GATC sites only, or "GATC genome" (Figure 4.1C). We demarcated this "GATC genome" into "chromosomes" whenever there was a region over 5000bp that lacked GATC sites. We then used Scripture to identify enriched peaks, and mapped these peaks back to the real genome.

We compared these Dam enriched peaks to the K562 chromatin state annotations, normalized by GATC sites. We found that the Dam protein has roughly 4-fold depletion in Polycomb regions, and nearly 3-fold enrichment in repetitive regions or regions subject to copy-number variation (Figure 4.1D). Bias existed at other genomic regions, but was not as pronounced. Thus, we must normalize any DamID-seq maps for chromatin proteins by the Dam protein.

To account for both the uneven distribution of GATC sites in the genome and the inherent bias of the Dam protein, we developed a peak caller that uses the log2 of the ratio between the protein over the Dam alone (Figure 4.1E). We called peaks that were above a certain quantile for each protein, based on known literature. Visual inspection confirmed that these peaks correlate with enriched regions in the protein track, except when there is high enrichment in the Dam control track.

We validated that the log2 peak caller accounts for the inherent GATC and GC bias in our method, as shown in Chapter 3. For 2kb bins across one representative chromosome, we compared the number of GATCs with the log2 peak caller reads by plotting a log-scale density scatter plot (Figure 4.1F). We found that the correlation coefficient was reduced to 0.054, indicating that the number of GATC sites no longer biases our read numbers when we process with the log2 peak caller. Similarly, when we compare the G+C percentage with the log2 peak caller reads (Figure 4.1G), we found that the correlation coefficient was reduced to -0.022, indicating that the G+C percentage no longer biases our read numbers.

Validation of DamID-seq by Comparison with ChIP-seq

We mapped the binding of three chromatin proteins by both DamID-seq and ChIP-seq in K562 cells: EZH2, RNF2, and CBX8. Since ChIP-seq is an established method for mapping proteins, we used it to validate our DamID-seq method and peak caller.

By visual inspection, the DamID-seq tracks for EZH2, RNF2 and CBX8 correlate fairly well with their respective ChIP-seq tracks (Figure 4.2A). Many of the peaks seem to overlap between the two methods.

Figure 4.2. Validation of DamID-seq by Comparison with ChIP-seq

(A) DamID-seq tracks and peaks, as called by the log2 peak caller, and ChIP-seq tracks and peaks, as called by Scripture, for EZH2, RNF2, and CBX8. (B) Venn diagram of 2 kb windows, excluding windows with less than 2 or greater than 10 GATC sites and accounting for adjacent windows, according to overlap with DamID-seq peaks (purple), ChIP-seq peaks (peach), or both (overlap), for EZH2, RNF2, and CBX8. The area of overlap is proportional to the percentage of windows in each class. (C) Classification of ChIP-seq peaks for EZH2, RNF2, and CBX8 according to whether they 1) overlap with DamID-seq peaks (blue), 2) are within 5 kb of a DamID-seq peak (red), or 3) are greater than 5kb of a DamID-seq and therefore "undetected" by DamID-seq (green). Numbers under the bars represent the total number of peaks called. (D) Classification of DamID-seq peaks for EZH2, RNF2, and CBX8 according to whether they 1) overlap with ChIP-seq peaks (blue), 2) are within 5 kb of a ChIP-seq peak (red), or 3) are greater than 5kb of a ChIP-seq and therefore "novel" (green). Numbers under the bars represent the total number of peaks called. (E) Percentage of peaks overlapping H3K27me3 for EZH2, RNF2, and CBX8 according to whether they were detected by 1) ChIP-seq only (green), 2) both ChIP-seq and DamID-seq (blue), or 3) DamID-seq only (red).

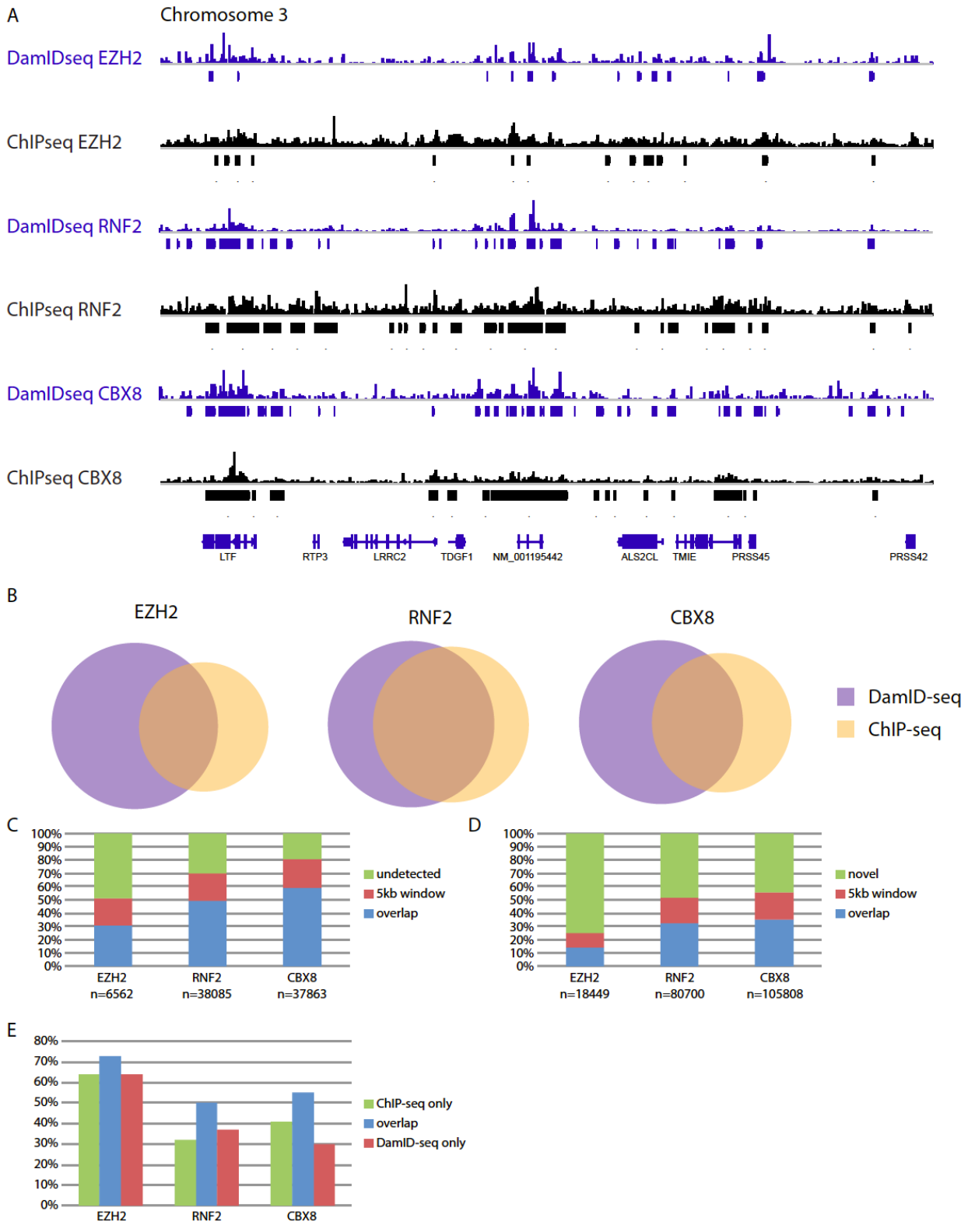

Figure 4.2 (Continued).

Some of the called peaks have different lengths or slightly shifted locations, which may be due to the uneven distribution of GATC sites. Overall, there appears to be consistency between the two methods.

To quantify the correlation between DamID-seq and ChIP-seq systematically across the genome, we divided the genome into 2 kb windows, and identified windows that overlapped with DamID-seq peaks, ChIP-seq peaks, or both (Figure 4.2B). We excluded windows that contained less than 2 or greater than 10 GATC sites, as that is the range of detection for DamID-seq. We also counted adjacent windows in the overlap percentage. For EZH2, we found that 56% of 13876 ChIP-seq windows, and 33% of 20847 Dam-seq windows, overlapped windows of the opposite technology. For RNF2, the overlap contained 73% of 96232 ChIP-seq windows and 63% of 100272 DamID-seq windows, and for CBX8, 71% of 63757 ChIP-seq windows and 44% of 103596 DamID-seq windows. Thus, both methods have respectable overlap on a 2 kb scale.

To examine whether the overlap between the ChIP-seq and DamID-seq windows are statistically significant, we performed a hypergeometric test. We found that for EZH2, the overlap between the two methods is significant, with a p-value of 2.57×10^{-9}. Similarly, for RNF2 and CBX8, the p-values were both less than 1×10^{-9}. Therefore, the overlap between the peaks called by the two technologies on a 2kb scale is statistically significant.

Since neither DamID-seq nor ChIP-seq peaks consistently correlate with 2 kb windows, we then examined the overlap on the basis of each ChIP-seq peak (Figure 4.2C). Of the 6562 peaks called in EZH2 by ChIP-seq, about 31% overlap with DamID-seq peaks, and an additional 20% are within 5 kb of a DamID-seq peak. Thus, about 51% of the ChIP-seq data can be accounted for by DamID-seq. The overlap is greater for RNF2 and CBX8. Of the 38085 peaks called in RNF2 by ChIP-seq, roughly 50% directly overlap, and an additional 20% are within

5kb, of DamID-seq peaks. Of the 37863 CBX8 peaks called by ChIP-seq, roughly 59% directly overlap, and an additional 22% are within 5kb, of DamID-seq peaks. Thus, over 70% and over 80% of ChIP-seq peaks can be accounted for by DamID-seq in RNF2 and CBX8, respectively.

This suggests that DamID-seq can account for more ChIP-seq peaks when performed on proteins that cover a greater percentage of genome. The mean ChIP-seq peak length for EZH2, RNF2, and CBX8 is 3116bp, 3659bp, and 4586bp, respectively. Thus, it also appears that DamID-seq can better account for ChIP-seq peaks if they are longer in length. This makes sense given that DamID-seq has a lower resolution compared to ChIP-seq, and depends on the distribution of GATC sites. Therefore, we recommend using DamID-seq on proteins with expected broad binding patterns.

We next examined whether DamID-seq identifies novel peaks that are not found by ChIP-seq (Figure 4.2D). Of the 18449 peaks called in EZH2 by DamID-seq, about 14% overlap with ChIP-seq peaks, and an additional 11% are within 5kb of a ChIP-seq peak. Therefore, 75% of DamID-seq peaks for EZH2 in K562 cells are novel. For RNF2, of the 80700 DamID-seq peaks, about 33% overlap with ChIP-seq peaks, an additional 19% are within 5kb, and 48% are novel. For CBX8, of the 105808 DamID-seq peaks, 35% overlap, and another 20% are within 5kb of, ChIP-seq peaks, while 44% are novel.

We then asked whether the novel DamID-seq peaks were false positives or not. We used the H3K27me3 histone modification state as a proxy for true binding peaks, since it is known that Polycomb catalyzes this mark (Kerppola 2009). We examined the percentage overlap with H3K27me3 for peaks detected by ChIP-seq only, DamID-seq only, or both methods. For EZH2, we found that while 73% of peaks detected by both methods overlapped with H3K27me3, 64% of peaks detected solely by DamID-seq overlapped this mark. This percentage is the same as that

for peaks detected solely by ChIP-seq. This suggests that over 87% of novel DamID-seq peaks are real, and that DamID-seq has a similar accuracy as ChIP-seq. For RNF2, 50% of peaks detected by both methods overlapped H3K27me3, while 37% of DamID-seq peaks and 32% of ChIP-seq peaks overlapped this mark. Here, at least 74% of novel DamID-seq peaks appear accurate; in this case, DamID-seq may be even more reliable than ChIP-seq. Finally, for CBX8, 55% of peaks detected by both methods, 30% of peaks detected by DamID-seq only, and 41% of peaks detected by ChIP-seq only, overlapped with H3K27me3. Of the three proteins we compared, the accuracy of novel DamID-seq peaks for CBX8 appears the least accurate (over 54%). However, this protein's peaks had the most overlap between the two methods (over 80%), so it seems reasonable that the few outlying novel peaks may be more likely to be false positives compared to the other two proteins. Overall, it appears that the large majority of novel DamID-seq peaks are real, and the accuracy of our method is similar to ChIP-seq, for all proteins we examined.

There are several potential explanations for the novel peaks identified by DamID-seq. First of all, since DamID-seq relies on enzyme activity over 3 days, while ChIP-seq relies on formaldehyde crosslinking within 10 minutes, DamID-seq may be able to identify additional domains of enrichment. For instance, a Dam-fused protein that is loosely or transiently bound to its targets can still methylate adenines at nearby GATC sites, while formaldehyde may be unable to crosslink these regions within a 10-minute timeframe. Secondly, since DamID-seq is based on a fusion protein, rather than an antibody as in ChIP-seq, it is possible that it is able to detect binding in areas that may be inaccessible to antibodies, such as regions of high compaction. Thirdly, DamID-seq is subject to potential artifacts resulting from the protein's fusion to Dam. This exogenous fusion protein may have binding sites that are neither reflected in the

endogenous protein nor accounted for by the Dam only control. It may also compete with the endogenous protein to skew the canonical function of the protein in the cell. Finally, the DamID-seq peaks were called by the log2 peak caller, wile ChIP-seq peaks were called by Scripture. An independent, positive control for the protein's true binding sites, additional sequencing reads for the DamID-seq maps, and further refinement of the statistical peak caller may help distinguish between these different explanations.

DamID-seq Maps Genomewide Binding of 12 Chromatin Proteins

We used DamID-seq to map the genomewide binding of EZH2, BMI1, RNF2, CBX1, CBX2, CBX3, CBX5, CBX6, CBX7, CBX8, UHRF1, and LMNB1 (Figure 4.3A). For comparison, we also included the ChIP-seq maps of histone modifications H3K27me3 and H3K9me3. We used our log2 peak caller to identify enriched peaks for each protein. The percent of genome covered by peaks for each protein was roughly 5%, except for EZH2 (0.7%), BMI1 (0.4%), and LMNB1 (8.3%).

By visual inspection, we found that the DamID-seq tracks for these proteins tended to look similar to either H3K27me3 or H3K9me3. CBX1 and CBX2 seemed to have qualities of both histone modification patterns.

To begin to study the correlation between different proteins more systematically, we made scatter plots comparing two proteins at a time using 2kb bins across the genome (Figure 4.3B).

Figure 4.3. DamID-seq Maps of 12 Chromatin Proteins

(A) DamID-seq tracks of Polycomb proteins EZH2, BMI1, RNF2, CBX6, CBX7, CBX8 (blue), Polycomb protein CBX2 and heterochromatin protein CBX1 (light blue), heterochromatin proteins CBX3, CBX5, UHRF1, LMNB1 (red), and Dam only control (black). Interspersed are ChIP-seq tracks of histone modifications H3K27me3 (green) and H3K9me3 (purple). (B) Scatter plots comparing 2 kb bins of reads for indicated DamID-seq maps. The red number in the upper left corner is the correlation.

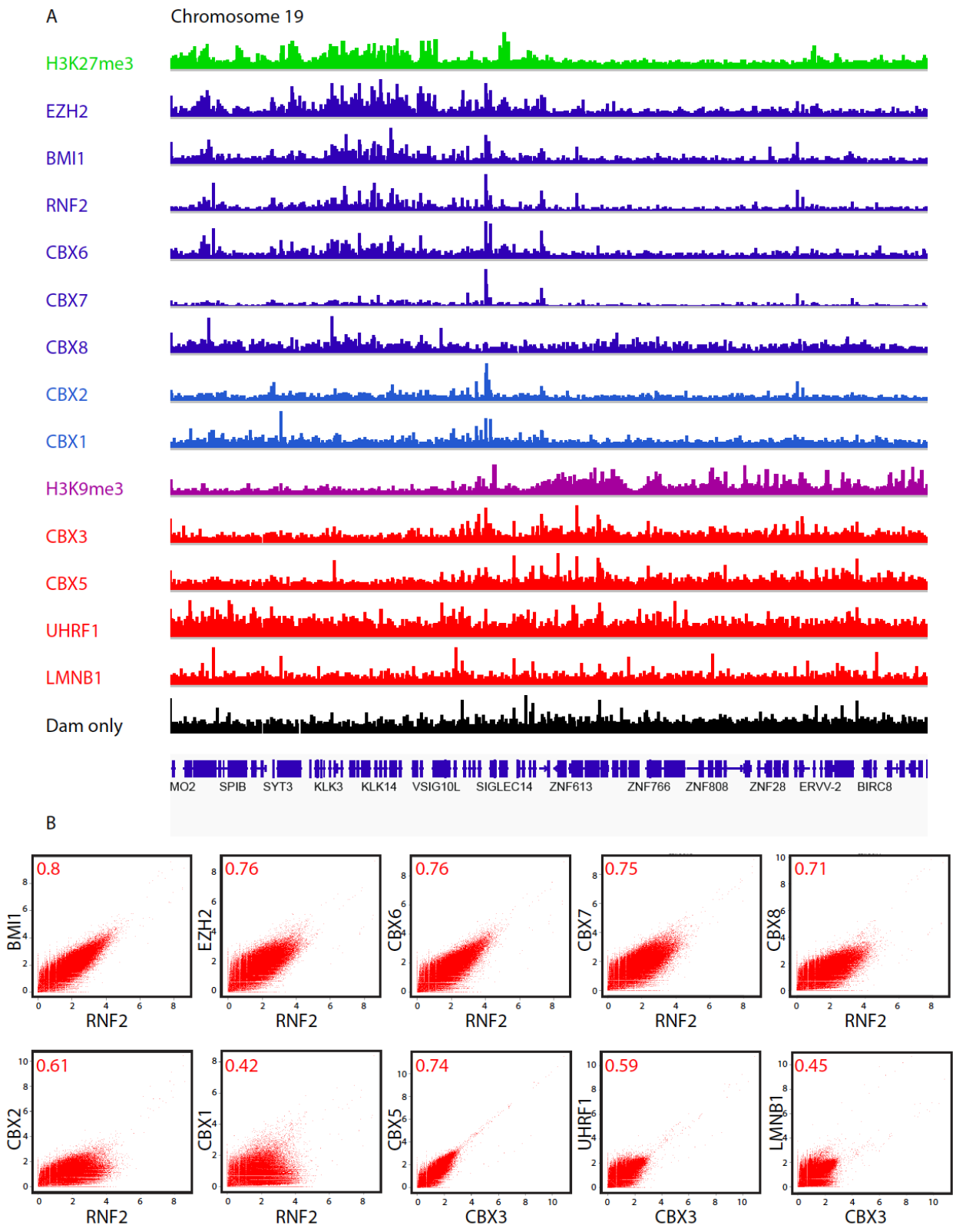

Figure 4.3 (Continued).

We found the highest correlations between members of the Polycomb group: RNF2 had a correlation of 0.8 with BMI1, 0.76 with EZH2, 0.76 with CBX6, 0.75 with CBX7, and 0.71 with CBX8. The correlation between RNF2 and CBX2 and CBX1 was much less: 0.61 and 0.42, respectively. CBX3 and CBX5, both dHP1 homologs, had a strong correlation: 0.74. The correlation between CBX3 and UHRF1 and LMNB1 was much less: 0.59 and 0.45, respectively. Again, it appears that CBX2 and CBX1 have unexpected binding patterns, given their homology to *d*Pc and *d*HP1, respectively.

Chromatin Proteins Cluster into Two Major Modules

To systematically compare the 12 chromatin proteins with each other, we present a matrix of correlations between each protein and the histone modifications, H3K27me3 and H3K9me3 (Figure 4.4A). We find that our set of chromatin proteins cluster into two major modules: 1) those that are members of Polycomb complexes, and 2) those that bind to heterochromatin. Surprisingly, CBX2, which is a homolog of *d*Pc, clustered with heterochromatin proteins.

The Polycomb cluster proteins (EZH2, CBX6, CBX7, CBX8, RNF2, and BMI1) have strikingly high correlations with each other (Figure 4.4A). They are also all positively correlated with H3K27me3. This is expected since EZH2 is the catalytic subunit of Polycomb repressive complex 2 (PRC2), which methylates H3K27. Additionally, RNF2, BMI1, CBX6, CBX7, and CBX8 are all members of Polycomb repressive complex 1 (PRC1), which can recognize H3K27me3. It appears that in K562 cells, the targets of these two complexes are quite similar.

Figure 4.4. Chromatin Proteins Cluster into Two Major Modules

(A) Matrix of correlations between the 12 DamID-seq protein maps, ChIP-seq of H3K27me3, and ChIP-seq of H3K9me3. The black boxes demarcate the Polycomb cluster and the heterochromatin cluster. (B) Pie charts for Polycomb cluster proteins EZH2, BMI1, RNF2, CBX6, CBX7, and CBX8 showing the percentage of peaks in six aggregated state annotations: promoters (orange), transcribed regions (blue), distal elements (black), Polycomb (yellow), heterochromatin (red), and repetitive/copy-number variants (green). State annotations are based on nine histone modification maps in human K562 cells (Ernst, Kheradpour et al. 2011). (C) Pie charts for heterochromatin cluster proteins CBX1, CBX3, CBX5, UHRF1, LMNB1, and CBX2 showing the percentage of peaks in six aggregated state annotations: promoters (orange), transcribed regions (blue), distal elements (black), Polycomb (yellow), heterochromatin (red), and repetitive/copy-number variants (green). State annotations are based on nine histone modification maps in human K562 cells (Ernst, Kheradpour et al. 2011).

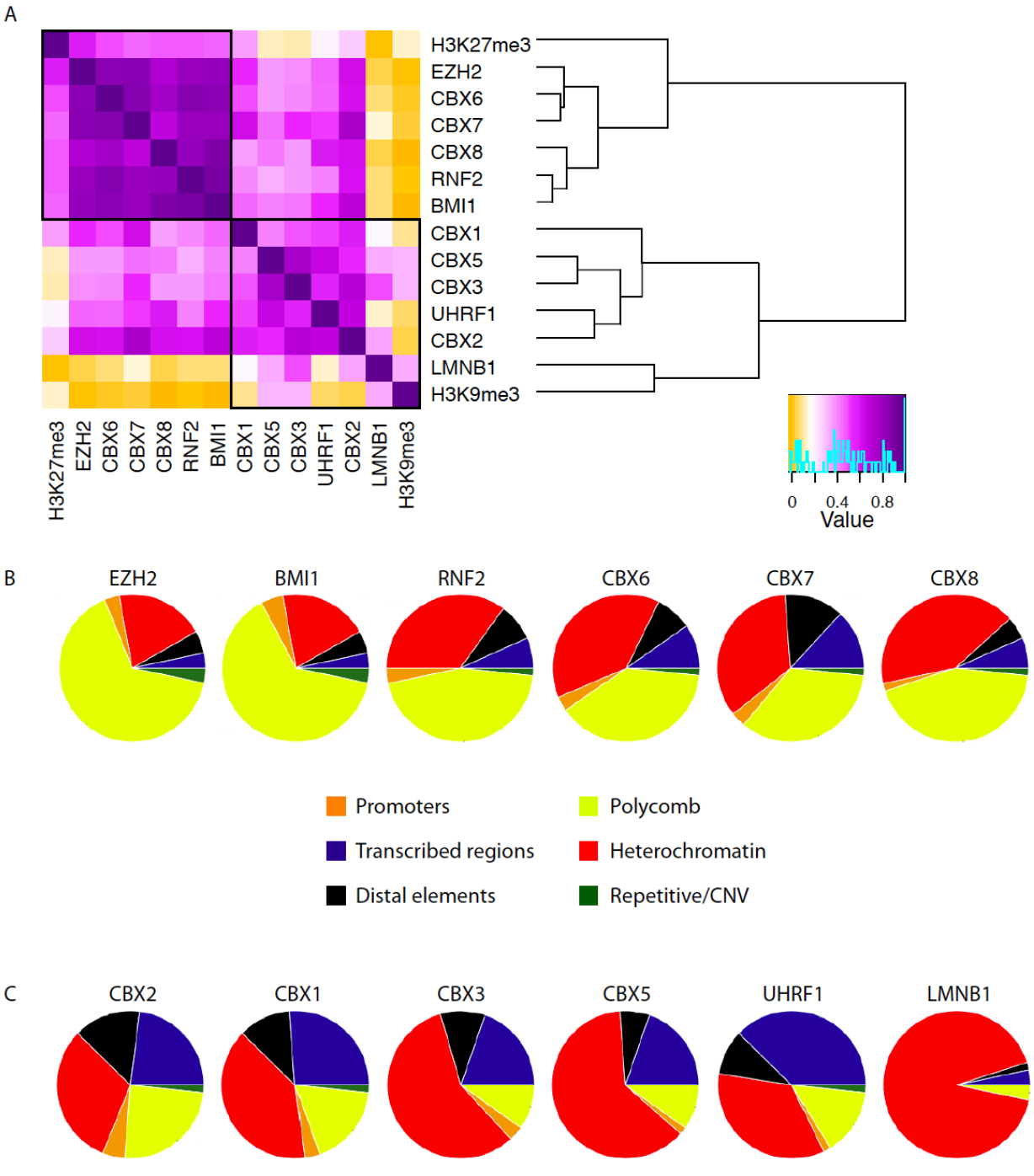

Figure 4.4 (Continued).

The heterochromatin cluster proteins (CBX1, CBX5, CBX3, UHRF1, CBX2, and LMNB1) have greater differences between each other (Figure 4.4A). Interestingly, CBX3, CBX5, and LMNB1 are positively correlated with H3K9me3, while CBX1, CBX2, and UHRF1 are not. As CBX1, CBX3, and CBX5 are all homologs of *d*HP1, it appears that CBX1 has diversified from the other two. UHRF1 has been previously reported to bind H3K9me3 (Rush, Cho et al. 2002), but the correlation seems to be weak in K562 cells. CBX2 surprisingly correlated best with UHRF1, and has clearly diverged in function from the other CBX proteins that are *d*Pc homologs.

To characterize the binding targets of the 12 chromatin proteins, we compared the DamID-seq peaks of each with genomewide chromatin state annotations for K562 cells (Ernst, Kheradpour et al. 2011), which we aggregated into "promoters," "transcribed regions," "distal elements," "Polycomb," "heterochromatin," and "repetitive/copy-number variants." These state annotations are based on ChIP-seq maps of histone modifications, rather than chromatin proteins.

For the Polycomb cluster, we found that all proteins bound a large proportion of the "Polycomb state" (Figure 4.4B). While only 9% of the genome is in the "Polycomb state," 66% of EZH2 and 64% of BMI1 peaks are in this state. These values constitute 7.3- and 7.1-fold enrichments, respectively. The other Polycomb cluster proteins also show great enrichment for this state. Specifically, 44% of RNF2 peaks, 39% of CBX6 peaks, 34% of CBX7 peaks, and 43% of CBX8 peaks are in the "Polycomb state." These values are all at least 3.8-fold enriched over background. Thus, the Polycomb proteins we mapped with DamID-seq seem to be going to expected locations as defined by ChIP-seq of histone modifications.

For the heterochromatin cluster, we found that all proteins bound a large proportion of the "heterochromatin state" (Figure 4.4C). Notably, 58% CBX3, 61% of CBX5, and 92% of

LMNB1 targets are in this state, while only 19% of the genome is in this state. These values constitute a 3.0-, 3.2-, and 4.8-fold enrichment over background. CBX1, CBX2, and UHRF1 all had modest enrichment for the heterochromatin state, and had minor enrichment for the Polycomb state as well. Specifically, 39% of CBX1 peaks, 32% of CBX2 peaks, and 35% of UHRF1 peaks were in the "heterochromatin state." The "Polycomb state" accounted for 18%, 24%, and 14% of these protein's peaks, respectively. This translates to at least a 1.7-fold enrichment of the "heterochromatin state" and 1.6-fold enrichment of the "Polycomb state" for these three proteins. This again highlights the diversification of CBX1 and CBX2 from their homologous proteins.

Polycomb Cluster: EZH2, BMI1, RNF2, CBX6/7/8 Bind Developmental Genes

To explore the functional significance of the binding patterns of these chromatin proteins, we identified the genes located under the binding peaks for each of the DamID-seq protein maps, and performed Gene Ontology enrichment analysis (Table 4.1). We found great redundancy in the gene targets of the Polycomb proteins EZH2, BMI1, RNF2, CBX6, CBX7, and CBX8, and examined their gene targets as a whole.

We found that the Polycomb proteins bind genes involved in development and localization with a p-value of less than $10x10^{-7}$ (Table 4.1). Examples of these genes include the *HOXC* cluster, *NEUROD2*, and *GNAS* (Figure 4.5A). The Hox locus contains clusters of genes and non-coding RNA involved in embryonic development (Rinn, Kertesz et al. 2007; Tschopp and Duboule 2011).

Polycomb	p-value	-log10 p-value
developmental process	2.90E-08	7.54
localization	9.55E-07	6.02
multicellular organismal development	1.78E-06	5.75
transport	1.15E-05	4.94
establishment of localization	1.41E-05	4.85
Heterochromatin	**p-value**	**-log10 p-value**
cytoplasm	0.00242	2.62
protein binding	0.0126	1.90
hydrolase activity	0.0303	1.52
CBX1, 3, 5	**p-value**	**-log10 p-value**
sensory perception of smell	0	∞
sensory perception of chemical stimulus	0	∞
olfactory receptor activity	0	∞
ion binding	3.42E-05	4.47
metal ion binding	4.83E-05	4.32
UHRF1	**p-value**	**-log10 p-value**
sensory perception of smell	0	∞
rhodopsin-like receptor activity	0	∞
olfactory receptor activity	0	∞
sensory perception of chemical stimulus	0	∞
cytoplasm	2.11E-24	23.7
LMNB1	**p-value**	**-log10 p-value**
membrane	1.14E-16	15.9
membrane part	1.44E-14	13.8
olfactory receptor activity	1.51E-13	12.8
plasma membrane	1.27E-12	11.9
sensory perception of smell	5.26E-12	11.3
CBX2	**p-value**	**-log10 p-value**
macromolecule modification	6.12E-05	4.21
post-translational protein modification	0.00011	3.96
protein modification process	0.00011	3.96
adenyl ribonucleotide binding	0.00011	3.96
ATP binding	0.00011	3.96

Table 4.1. Gene Ontology Enrichment Analysis

Enriched GO terms for gene targets of Polycomb cluster proteins (EZH2, BMI1, RNF2, CBX6, CBX7, CBX8), heterochromatin cluster proteins (CBX1, CBX3, CBX5, UHRF1, LMNB1), CBX1/3/5, UHRF1, LMNB1, and CBX2. The Benjamini-corrected p-value and the -log10 p-value are shown.

Figure 4.5. Gene Targets of Polycomb and Heterochromatin Proteins

(A) DamID-seq maps of Polycomb proteins EZH2, BMI1, RNF2, CBX6, CBX7, and CBX8 at the HOX cluster containing *HOXC13*, *HOXC12*, lincRNA HOTAIR, and *HOXC11*; and genes *NEUROD2* and *GNAS*. (B) DamID-seq maps of heterochromatin proteins CBX1, CBX3, CBX5, and UHRF1 at OR cluster containing *OR7D2*, *OR7D4*, and *OR7E24*; gene *SCNN1G*; and ZNF cluster containing *ZNF223*, *ZNF284*, *ZNF224*, and *ZNF225*. (C) DamID-seq maps of CBX2 at genes *DOLPP1*, *ZER1*, and *QTRT1*.

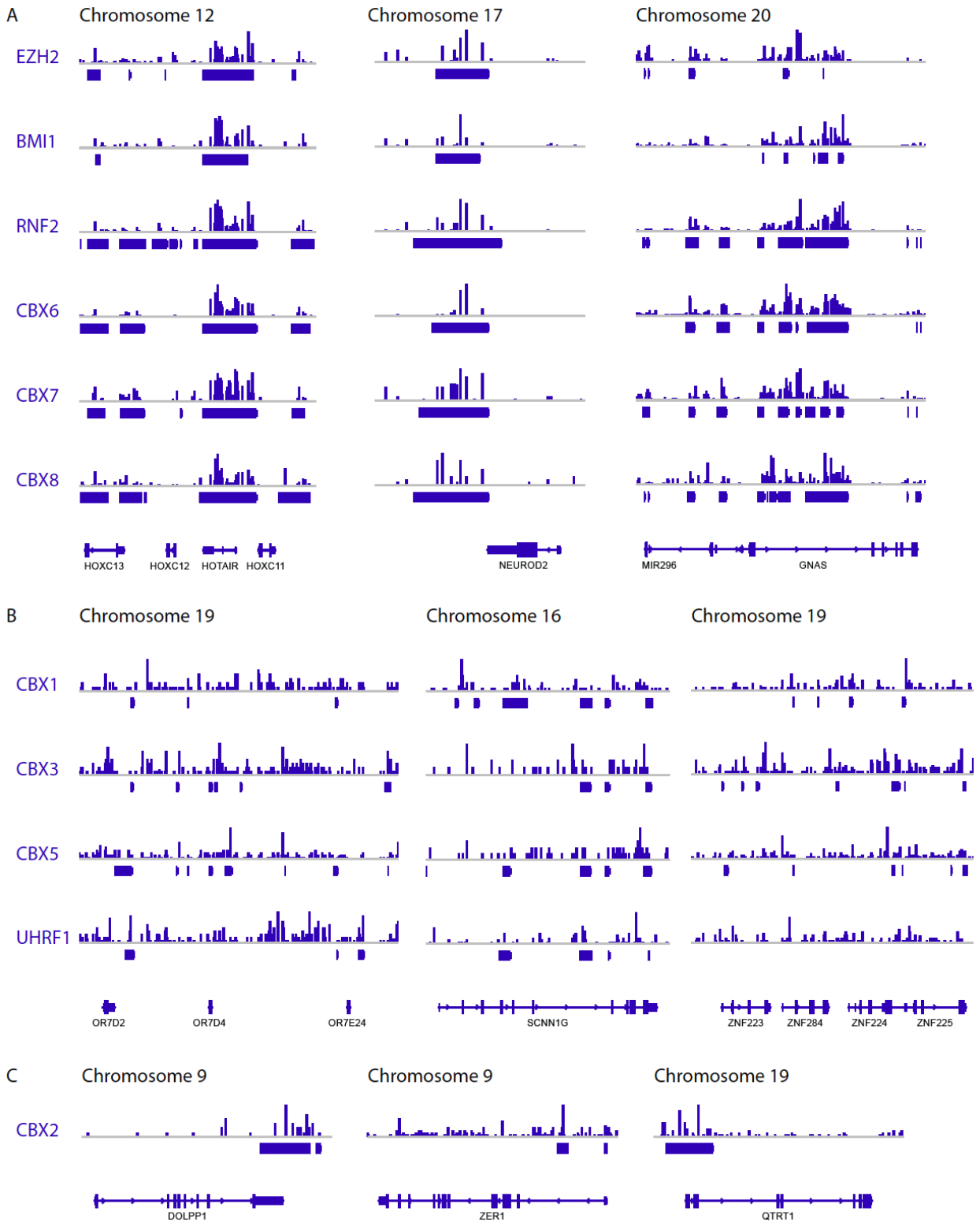

Figure 4.5 (Continued).

NEUROD2 is thought to determine and maintain the neuronal lineage during cell differentiation, and GNAS is known for its complex imprinting patterns (Messmer, Shen et al. 2012; Robson, Eaton et al. 2012). These results are consistent with previous reports that Polycomb is involved in repressing and poising developmental genes (Zhou, Goren et al. 2011).

We note that the Polycomb proteins exhibit relatively sharp binding peaks, rather than broad domains (Figure 4.3; Figure 4.5A). In the three above examples, the Polycomb proteins bind with strong peaks at the promoter, across the body, or at the 3'UTR of its targets. This may reflect the tight regulation of Polycomb localization.

Heterochromatin Cluster: CBX1/3/5 Bind Broad OR and ZNF Domains

As a whole, the gene targets of the heterochromatin proteins (CBX1, CBX3, CBX5, UHRF1, LMNB1) did not give remarkable GO enrichments (Table 4.1). The few significant terms, such as cytoplasm and protein binding, were very general and of low overlap. It appears that the heterochromatin proteins we mapped have different target genes. Thus, we examined the GO enrichments for genes of these proteins separately.

CBX1, 3, and 5 all bind a common pool of gene targets, despite the seeming diversity of CBX1 from the other two. These three *d*HP1 homologs bind genes involved in sensory functions; this is significant with a p-value of less than 10^{-24} (Table 4.1). For instance, they coat the length of the OR gene cluster containing *OR7D2*, *OR7D4*, and *OR7E24*, as well as *SCNN1G* (Figure 4.5B). Olfactory receptors comprise a large family of genes that initiate a neuronal response that triggers smell, and SCNN1G is a subunit of a non-voltage-gated sodium channel (Malnic, Godfrey et al. 2004). In addition, CBX1/3/5 targets genes involved in ion and metal ion

binding (Table 4.1). For instance, they bind the zinc finger (ZNF) cluster that includes *ZNF223*, *ZNF284*, *ZNF224*, and *ZNF225*. ZNF genes encode a large family of zinc finger proteins that bind zinc ions and are involved in transcriptional gene regulation (Lorenz, Dietmann et al. 2010). This finding is consistent with maps that show that CBX1 binds KRAB-ZNF genes in MCF7 human breast carcinoma cell lines (Vogel, Guelen et al. 2006).

Like CBX1/3/5, UHRF1 also binds genes that are involved in sensory functions, such as OR clusters and *SCNN1G* (Table 4.1; Figure 4.5B). However, it does not seem to bind to ZNF clusters or genes involved in ion and metal ion binding.

LMNB1's gene enrichments are very general, such as membrane and membrane part (Table 4.1). Indeed, most of LMNB1 binding is in regions of the genome that are gene poor. Its binding domains are extremely broad, and confirm previous reports that LMNB1 defines Megabase domains (Guelen, Pagie et al. 2008).

It is notable that unlike Polycomb proteins, heterochromatin proteins bind broad domains that can be Megabases in length (Figure 4.3; Figure 4.5B). Both of CBX1/3/5's notable gene targets, OR and ZNF genes, are members of large gene families that occur in broad clusters. This may reflect the role of heterochromatin in structurally sequestering large expanses of genome.

CBX2 Binds Unique Gene Targets

Since CBX2 seems to have diversified binding patterns compared to the other Polycomb proteins, we examined whether these differences are reflected in its gene targets. We found that indeed, CBX2 has unique gene targets beyond those of the other Polycomb proteins.

Specifically, CBX2 is enriched at genes that are involved with modifying proteins and

binding adenyl ribonucleotide or ATP with a p-value of less than 0.00011 (Table 4.1). Examples of these genes include *DOLPP1*, *ZER1*, and *QTRT1* (Figure 4.5C). Dolichyl pyrophosphate phosphatase 1 (DOLPP1) is an enzyme that is thought to play a role in recycling dolichyl pyrophosphate during reactions in the endoplasmic reticulum (Rush, Cho et al. 2002). ZER1 is a little studied protein that is thought to help recruit the E3 ubiquitin ligase complex ZER1-CUL2-Elongin BC to its targets (Kim, Bennett et al. 2011). As a third example, queuine tRNA-ribosyltransferase 1 (QTRT1) is the catalytic subunit of RNA-guanine transglycosylase, which modifies tRNAs (Chen, Brooks et al. 2011). Thus, CBX2 may play a unique role as a PRC1 component involved in cellular enzymatic reactions.

We note that CBX2, like the other Polycomb proteins, exhibits punctate binding peaks, rather than broad domains (Figure 4.3; Figure 4.5C). In the three above examples, CBX2 binds at the 3' UTR of *DOLPP1*, the first intron of *ZER1*, and the promoter of *QTRT1*. Thus, it also appears that CBX2 can bind at any location in its target genes. Knockdown studies of CBX2 in mammalian cells are needed to address whether it can functionally repress genes regardless of binding location.

DISCUSSION

Here, we presented the application of our DamID-seq technology to map the genomewide binding of 12 chromatin proteins in human K562 cells, and offered a novel peak caller for identifying binding domains from such data. We envision that DamID-seq, which is able to map broadly and loosely bound chromatin proteins, will complement ChIP-seq, which is better suited for mapping histone modifications and transcription factors, to comprehensively characterize

chromatin in mammalian cells.

In general, ChIP-seq may be a better choice for proteins that it can readily map, such as transcription factors and DNA-binding proteins. ChIP-seq measures binding within 10 minutes and has a resolution of 25bp, while DamID-seq measures binding within 3 days and has a resolution of 1kb or greater. Furthermore, ChIP-seq measures the endogenous protein, while DamID-seq measures an exogenous fusion protein, though it is lowly expressed. At this time, ChIP-seq has also been developed for small cell numbers and has been performed on as low as 10^4 cells (Adli, Zhu et al.); DamID-seq has only been performed on 10^6 cells, though no attempt has been made yet to find the lower limit for this technology.

However, ChIP-seq is unable to map proteins without high-quality antibodies, most loosely, transiently, or broadly bound proteins, proteins in regions of compact chromatin, and mutant proteins. In these cases, DamID-seq may be uniquely able to provide genomewide mapping data. This is because DamID-seq does not rely on antibodies, crosslinking, or the solubility of chromatin. Furthermore, DamID-seq takes a reduced representation of the genome; specifically, it only reads GATC sites. Thus, it can map broadly distributed proteins with fewer sequencing reads. Finally, DamID-seq is not subject to crosslinking artifacts or antibody cross-reactivity, which affects ChIP-seq experiments.

Our DamID-seq maps of 12 chromatin proteins reveal two major modules: 1) Polycomb-related, and 2) heterochromatin-related. The PRC2 component EZH2, and the PRC1 components BMI1, RNF2, CBX6, CBX7, and CBX8 all bind a common set of developmental gene targets that is consistent with previous reports (Zhou, Goren et al. 2011). In contrast, heterochromatin proteins CBX1/3/5 bind OR and ZNF genes, which both occur in broad clusters in the genome (Malnic, Godfrey et al. 2004; Lorenz, Dietmann et al. 2010).

Surprisingly, our maps show that the PRC1 component CBX2 exhibits unique binding to genes involved with modifying proteins, and may play a role in regulating cellular enzymatic reactions. CBX2 has been implicated in the repression of ovarian development in XY gonads by regulation of SRY, and in the maintenance of chromosome stability in the mammalian germline (Baumann and De La Fuente 2011; Yap and Zhou 2011). Interestingly, it is the only CBX protein that does not co-elute with RING1 in tandem affinity purification studies (Vandamme, Volkel et al. 2011). However, the role of CBX2 in mammalian somatic cells is yet to be studied. Our analyses suggest that knockdown of CBX2 in K562 cells may elucidate whether it plays an additional role in critical cellular enzymatic reactions in somatic cells.

Since the structure of CBX2 is so similar to other CBX proteins, one may question why it would exhibit unique binding patterns (Yap and Zhou 2011). We note that CBX1 and CBX5 have been found to have varying binding patterns depending on humoral signals and microenvironmental cues (Ritou, Bai et al. 2007), and such factors may also affect the binding of CBX2. Additionally, the chromodomain of CBX2 may be subject to post-translational modifications and nucleic acid binding, which would also alter the protein's binding patterns (Yap and Zhou 2011). Furthermore, temporal changes, such as those due to the cell cycle and cell differentiation, add additional complexity to interpreting the DamID-seq maps, which reveal composite binding over three days. Thus, additional mechanistic studies are needed to reveal the conditions in which CBX2 displays unique binding.

Our findings support a model in which the genome is organized into large domains, with Polycomb proteins repressing developmental genes at bodies, and heterochromatin proteins sequestering broad OR and ZNF clusters near the nuclear lamina. The Polycomb and heterochromatin domains appear to exhibit distinct properties. While Polycomb domains appear

more punctate, and may be more susceptible to local binding changes (as in the case of CBX2), heterochromatin domains appear broader, and may be more stable (as in the case of CBX1/3/5). These findings may be just the beginning of an appreciation of the distinct properties of higher-order domains.

METHODS

Plasmid Construction

We obtained plasmids pLgw V5-EcoDam (Dam only negative control), pLgw EcoDam-V5-RFC1 (N-terminus Dam vector), pLgw RFC1-V5-EcoDam (C-terminus Dam vector), and pLgw CBX1-V5-EcoDam (CBX1-Dam positive control) from the Bas van Steensel laboratory. We obtained ORFs in plasmid pDONR221 (or pDONR201) for CBX2 (cloneID HsCD00080034), CBX3 (cloneID HsCD00296031), CBX5 (cloneID HsCD00079893), CBX6 (cloneID HsCD00045684), CBX7 (HsCD00079712), CBX8 (cloneID HsCD00079972), EZH2 (cloneID HsCD00039865), BMI1 (cloneID HsCD00000297), RNF2 (cloneID HsCD00044984), UHRF1 (cloneID HsCD00079664) and LMNB1 (clone ID HsCD00043675) from the PlasmID collection at the Dana-Farber/Harvard Cancer Center DNA Resource Core.

We used Invitrogen Gateway® Cloning technology to clone these ORFs into either pLgw EcoDam-V5-RFC1 (N-terminus Dam vector) or pLgw RFC1-V5-EcoDam (C-terminus Dam vector), depending on whether the available ORF had a stop codon. Namely, these proteins were cloned with an N-terminus Dam: CBX3, CBX6, BMI1, RNF2, and LMNB1. These proteins were cloned with a C-terminus Dam: CBX1, CBX2, CBX5, CBX7, CBX8, EZH2, and UHRF1.

Cell Culture

293T cells were grown according to standard protocols in Gibco KO DMEM media supplemented with 10% fetal bovine serum (FBS, Atlas Biologicals, F-0500-A), 1% Penicillin/Streptomycin (Invitrogen, 15140122), and 1% Glutamax. K562 erythrocytic leukemia cells (ATCC CCL-243) were grown according to standard protocols in RPMI 1640 media (Invitrogen, 22400105) supplemented with 10% fetal bovine serum (FBS, Atlas Biologicals, F-0500-A) and 1% Penicillin/Streptomycin (Invitrogen, 15140122).

Lentiviral Production and Infection

293T cells were grown in 15cm dishes until 60-80% confluence. 1140uL of DMEM was combined with 60uL of Fugene. After 5 min, the following three plasmids were added: Gag, pol and rev plasmid (6ug), VSV envelope plasmid (3ug), and specific cloned Dam-protein plasmid or GFP (9ug). This mixture was incubated at room temperature for 5 min, then added dropwise to the 293T cells. After 8-12 hours, the media was replaced with 12mL fresh medium. After 72 hours, the virus was collected filtered through 0.45 uM. The virus was ultracentrifuged at 28000 rpm for 2 hours in an SW41Ti rotor at 4°C. The virus was resuspended in 100uL PBS, and left at 4°C overnight.

K562 cells were counted with a hemacytometer and 1.5 million cells were allocated per each infection. Each aliquot of cells was spun down and resuspended in 3mL fresh media, and plated in one well of a 6-well plate. 2uL Polybrene (10mg/mL) and 30uL of concentrated virus

was added. The cells were spin-infected at 2500 rpm, for 90 min, at room temperature. The cells were then returned to 37°C overnight.

The following day, the infected cells were spun down and resuspended in 3mL fresh media. After 48 hours later, the cells were harvested and the gDNA was isolated using the Qiagen DNA Micro Kit, "Isolation of gDNA from Small Volumes of Blood" protocol. The DNA was eluted in 200uL buffer AE, and quantified by Nanodrop.

DamID Library Preparation and Sequencing

The gDNA was ethanol precipitated, and dissolved in TE pH7.5 to a concentration of 1 ug/uL. 2.5uL of gDNA was digested with 0.5uL of DpnI (NEB, 20 U/uL) at 37°C overnight in PCR tubes. DpnI was inactivated by heating to 80°C for 20 min. The DpnI-digested gDNA was ligated with the adaptor AdR, which is made by mixing and slowly annealing AdR-top (5' CTAATACGACTCACTATAGGGCAGCGTGGTCGCGGCCGAGGA 3') and AdR-bottom (5' TCCTCGGCCG 3'). This ligation was completed using 1uL T4 Ligase (Roche, 5U/uL) for 2 hours at 16°C. The T4 ligase was inactivated by heating to 65°C for 10 min. The resulting 20uL volume reaction was diluted to 50uL with ddH2O.

To amplify the regions flanked by adaptors, the following PCR was setup: 10uL DNA, 5uL 10x cDNA PCR reaction buffer (Clontech), 0.625uL primer bio-Adr-PCR (5' bio-GGTCGCGGCCGAGGATC 3', 100uM), 1uL dNTPs (10mM), 1uL PCR advantage enzyme mix (Clontech, 50X), 32.375uL ddH2O. The PCR reaction program was as follows: 1 cycle of 68°C (10min), 94°C (1min), 65°C (5min), 68°C (15min); 3 cycles of 94°C (1min), 65°C (1min), 68°C (10min); 17 cycles of 94°C (1min), 65°C (1min), 68°C (2min). 5uL of the PCR products

were ran on a gel to verify successful digestion and amplification.

The PCR products were cleaned with the Qiagen MinElute PCR Purification kit, and eluted in 20uL ddH2O. Following quantification by Nanodrop, 3ug of each sample was diluted in 100uL ddH2O. These samples were sonicated with Covaris using the following settings: 10% duty cycle, 5 intensity, 200 cycles per burst, for 4.5 min total. 10uL of the Covaris-sonicated samples were ran on a gel to verify sonication to 100-500bp.

To pull down the biotinylated ends of the PCR products, we used Invitrogen Dynabeads® MyOne Streptavidin T1 beads. 50uL of these beads were washed three times with 50uL of 1X Binding and Washing (B&W) buffer (5mM Tris-HCl pH 7.5, 0.5mM EDTA, 1M NaCl). The washed beads were resuspended in 75uL of Covaris-sonicated DNA, 100uL of 2X B&W buffer, and 25uL H2O. This mixture was incubated at 4°C for 15min on a rotator. Following a quick spin and decanting the supernatant, the beads were washed three times with 200uL of 1X B&W buffer.

The bound DNA was removed off the beads by digestion with DpnII at 37°C for 1 hour. The supernatant was collected and cleaned using the Qiagen MinElute Reaction Cleanup kit. The resulting DNA was eluted in 20uL H2O and quantified with Qubit. qPCR analysis was performed to validate the DamID DNA before submission for sequencing.

Libraries of DamID samples were prepared according to the Illumina Genomic DNA protocol, as described previously (Mikkelsen et al., 2007). The DamID-seq libraries were sequenced on Illumina GAII sequencers according to standard Illumina protocols.

DamID-seq Data Analysis

Sequence reads were aligned to the human genome reference (hg19). We filtered out low-quality reads, and filtered in reads that map to GATC sites. The number of reads was counted at each GATC site. The reads were viewed using the Integrative Genomics Viewer (http://www.broadinstitute.org/igv/).

For our interim peak caller, these counts were translated to a "GATC genome," which was made by concatenating GATC sites only. This "GATC genome" was demarcated into "chromosomes" whenever there was a region over 5000bp that lacked GATC sites. Scripture was used to segment the Dam data into enriched peaks, and these peaks were translated back to the real genome.

For our log2 peak caller, we scanned the genome using a sliding window of 2500bp, with a 250bp step size, and computed a score as follows: log2((protein count + 1)/(Dam count +1)). The 250bp at the center of the window was marked as a peak if the score exceeded a given threshold. Thresholds were established for each protein based on known literature and respective ChIP-seq maps if available. Specifically, the following thresholds were used: 2.25 (CBX3, CBX5, LMNB1), 2.5 (CBX1, CBX2), 2.7 (UHRF1), 3.0 (CBX8), 3.5 (CBX6), 3.7 (CBX7), 3.8 (RNF2), 5.2 (EZH2), 5.7 (BMI1).

To construct the correlation matrix, we first assembled the "world of Dam peaks," or the total of all peaks from our DamID-seq K562 tracks. Then for every 2kb window in this foreground, we calculated the Pearson correlation between the counts for the given pair of proteins. We used hierarchical cluster analysis to assemble these correlations into the matrix.

We used GoSTAT (http://gostat.wehi.edu.au/) to find statistically overrepresented Gene Ontology terms for genes called by DamID-seq peaks. We used the goa_human GO gene-association database, a minimal length of 3 for considered GO paths, and corrected for multiple

hypothesis testing with Benjamini. We displayed the top 5 (or less) GO terms that had a p-value of less than 0.05.

Chromatin Immunoprecipitation Library Preparation and Sequencing

Cells were crosslinked in 1% formaldehyde for 10 min at 37°C, and then quenched with glycine for 5 min at 37°C. Fixed cells were lysed in 1% SDS, 10mM EDTA and 50mM Tris-HCl pH 8.1 supplemented with protease inhibitor (Roche), fragmented with a Branson Sonifier (model S-450D) at 4°C to a size range between 200 to 800bp, and precipitated by centrifugation. 5 to 10 ug of antibody was pre-bound by incubating with a mix of Protein-A and Protein-G Dynabeads (Invitrogen, 100-02D and100-07D, respectively) in blocking buffer (PBS supplemented with 0.5% TWEEN and 0.5% BSA) for 2 hr. Washed beads were added to the chromatin lysate, and then incubated overnight. Samples were washed 6 times with RIPA buffer, twice with RIPA buffer supplemented with 500 mM NaCl, twice with LiCl buffer (10 mM TE, 250mM LiCl, 0.5% NP-40, 0.5% DOC), twice with TE (10Mm Tris-HCl pH 8.0, 1mM EDTA), and then eluted in 0.5% SDS, 300 mM NaCl, 5 mM EDTA, 10 mM Tris Hcl pH 8.0 at 65°C. Eluate was incubated at 65°C overnight, and then treated with RNaseA (Roche) for 30 min and Proteinase K (NEB) for 2hr. DNA was purified using a Qiagen DNA purification kit.

Libraries of ChIP samples were prepared according to the Illumina Genomic DNA protocol, as described previously (Mikkelsen, Ku et al. 2007). The ChIP-seq libraries were sequenced on Illumina GAII sequencers according to standard Illumina protocols.

ACKNOWLEDGMENTS

We thank B. van Steensel for the DamID vectors; M. Suva for the lentivirus vectors and protocol; M. Garber for discussions on designing the peak caller; and J. Zhu for help with gene analyses. V.W.Z. was supported by an NSF Graduate Research Fellowship and National Defense Science and Engineering Graduate Fellowship. D.F. is advised and funded by the Jun S. Liu Laboratory. B.E.B. is a Charles E. Culpeper Medical Scholar and Early Career Scientist of the Howard Hughes Medical Institute. Research in the Bernstein Laboratory is supported by funds from the Burroughs Wellcome Fund, HHMI, and the NIH.

REFERENCES

Adli, M., J. Zhu, et al. "Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors." Nat Methods **7**(8): 615-618.

Baumann, C. and R. De La Fuente (2011). "Role of Polycomb Group Protein Cbx2/M33 in Meiosis Onset and Maintenance of Chromosome Stability in the Mammalian Germline." Genes (Basel) **2**(1): 59-80.

Chen, Y. C., A. F. Brooks, et al. (2011). "Evolution of eukaryal tRNA-guanine transglycosylase: insight gained from the heterocyclic substrate recognition by the wild-type and mutant human and Escherichia coli tRNA-guanine transglycosylases." Nucleic Acids Res **39**(7): 2834-2844.

Ernst, J., P. Kheradpour, et al. (2011). "Mapping and analysis of chromatin state dynamics in nine human cell types." Nature **473**(7345): 43-49.

Filion, G. J., J. G. van Bemmel, et al. (2010). "Systematic protein location mapping reveals five principal chromatin types in Drosophila cells." Cell **143**(2): 212-224.

Guelen, L., L. Pagie, et al. (2008). "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions." Nature **453**(7197): 948-951.

Guttman, M., M. Garber, et al. (2010). "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs." Nat Biotechnol **28**(5): 503-510.

Kaustov, L., H. Ouyang, et al. (2011). "Recognition and specificity determinants of the human cbx chromodomains." J Biol Chem **286**(1): 521-529.

Kerppola, T. K. (2009). "Polycomb group complexes--many combinations, many functions." Trends Cell Biol **19**(12): 692-704.

Kim, W., E. J. Bennett, et al. (2011). "Systematic and quantitative assessment of the ubiquitin-modified proteome." Mol Cell **44**(2): 325-340.

Lomberk, G., L. Wallrath, et al. (2006). "The Heterochromatin Protein 1 family." Genome Biol **7**(7): 228.

Lorenz, P., S. Dietmann, et al. (2010). "The ancient mammalian KRAB zinc finger gene cluster on human chromosome 8q24.3 illustrates principles of C2H2 zinc finger evolution associated with unique expression profiles in human tissues." BMC Genomics **11**: 206.

Malnic, B., P. A. Godfrey, et al. (2004). "The human olfactory receptor gene family." Proc Natl Acad Sci U S A **101**(8): 2584-2589.

Messmer, K., W. B. Shen, et al. (2012). "Induction of neural differentiation by the transcription factor NeuroD2." Int J Dev Neurosci **30**(2): 105-112.

Mikkelsen, T. S., M. Ku, et al. (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." Nature **448**(7153): 553-560.

Pauler, F. M., M. A. Sloane, et al. (2009). "H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome." Genome Res **19**(2): 221-233.

Ram, O., A. Goren, et al. (2011). "Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells." Cell **147**(7): 1628-1639.

Rinn, J. L., M. Kertesz, et al. (2007). "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs." Cell **129**(7): 1311-1323.

Ritou, E., M. Bai, et al. (2007). "Variant-specific patterns and humoral regulation of HP1 proteins in human cells and tissues." J Cell Sci **120**(Pt 19): 3425-3435.

Robson, J. E., S. A. Eaton, et al. (2012). "MicroRNAs 296 and 298 are imprinted and part of the GNAS/Gnas cluster and miR-296 targets IKBKE and Tmed9." RNA **18**(1): 135-144.

Rosnoblet, C., J. Vandamme, et al. (2011). "Analysis of the human HP1 interactome reveals novel binding partners." Biochem Biophys Res Commun **413**(2): 206-211.

Rush, J. S., S. K. Cho, et al. (2002). "Identification and characterization of a cDNA encoding a dolichyl pyrophosphate phosphatase located in the endoplasmic reticulum of mammalian cells." J Biol Chem **277**(47): 45226-45234.

Sexton, T., H. Schober, et al. (2007). "Gene regulation through nuclear organization." Nat Struct Mol Biol **14**(11): 1049-1055.

Tschopp, P. and D. Duboule (2011). "A genetic approach to the transcriptional regulation of Hox gene clusters." Annu Rev Genet **45**: 145-166.

Vandamme, J., P. Volkel, et al. (2011). "Interaction proteomics analysis of polycomb proteins defines distinct PRC1 complexes in mammalian cells." Mol Cell Proteomics **10**(4): M110 002642.

Vincenz, C. and T. K. Kerppola (2008). "Different polycomb group CBX family proteins associate with distinct regions of chromatin using nonhomologous protein sequences." Proc Natl Acad Sci U S A **105**(43): 16572-16577.

Vogel, M. J., L. Guelen, et al. (2006). "Human heterochromatin proteins form large domains containing KRAB-ZNF genes." Genome Res **16**(12): 1493-1504.

Vogel, M. J., D. Peric-Hupkes, et al. (2007). "Detection of in vivo protein-DNA interactions using DamID in mammalian cells." Nat Protoc **2**(6): 1467-1478.

Wen, B., H. Wu, et al. (2009). "Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells." Nat Genet **41**(2): 246-250.

Yap, K. L. and M. M. Zhou (2011). "Structure and mechanisms of lysine methylation recognition by the chromodomain in gene transcription." Biochemistry **50**(12): 1966-1980.

Zhou, V. W., A. Goren, et al. (2011). "Charting histone modifications and the functional organization of mammalian genomes." Nat Rev Genet **12**(1): 7-18.

Chapter 5:

Discussion

Discussion

Vicky W. Zhou[1,2,3,4]

1. Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA, 02114.

2. Center for Systems Biology and Center for Cancer Research, Massachusetts General Hospital, Boston, Massachusetts, USA, 02114.

3. Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 02142.

4. Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA, 02115.

"Chromatin" was coined by W. Flemming around 1880 to refer to "that substance in the cell nucleus which is readily stained," as seen under the microscope (Olins and Olins 2003). Since the advent of molecular biology, we have focused on studying chromatin one loci at a time. Now, armed with technology to scale the study of molecular features to a genomewide level, we are once again facing the task of modeling the whole genome in nuclear space.

SUMMARY AND SIGNIFICANCE

In this Thesis, we presented new frameworks for the study of mammalian chromatin organization (Chapter 1). These models highlight initial efforts to integrate genomewide information on histone modifications and chromatin proteins, and to merge these molecular insights with the three-dimensional organization of chromatin within the nucleus. We hope to spur development of new technologies to characterize higher-order chromatin structures with the goal of understanding genome regulation during differentiation, development, and disease.

We then advanced the understanding of Polycomb regulation by identifying GC-rich sequences that are necessary and sufficient for Polycomb repressive complex 2 (PRC2) recruitment in mammalian ES cells (Chapter 2). We also showed that the candidate transcription factor YY1 is not directly involved in PRC2 recruitment in these cells. Our findings on the determinants of Polycomb localization provide insight on the programming of gene expression in mammalian development.

Next, we adapted DamID for high-throughput sequencing, a method we called DamID-seq. This technology generates genomewide maps of chromatin proteins without crosslinking and antibodies, and thus is able to profile broadly, loosely, and transiently bound proteins

(Chapter 3). This method enables many previously unmappable proteins to be characterized in human cells for the first time.

We then used DamID-seq to map 12 chromodomain-containing and related proteins in K562 cells, and found that Polycomb proteins cluster together and bind developmental genes, while heterochromatin proteins cluster together and bind broad olfactory receptor and zinc finger domains (Chapter 4). We also identified CBX2 as a unique Polycomb protein that binds to genes involved with modifying proteins. These findings support the model of chromatin compartmentalization within the mammalian cell nucleus.

We note that in general, ChIP-seq may be a better choice for proteins that it can readily map, such as transcription factors and DNA-binding proteins, because it has a higher resolution, measures the endogenous protein, and has been performed on lower cell numbers. However, for the proteins that ChIP-seq is unable to map, including most chromatin proteins, DamID-seq may be uniquely able to provide genomewide mapping data. Furthermore, DamID-seq takes a reduced representation of the genome, thus requiring fewer sequencing reads to map broad proteins, and is not subject to crosslinking artifacts or antibody cross-reactivity (Chapter 4 Discussion).

Finally, we used chromosome conformation capture (3C) technology to reveal looping interactions between enhancers and promoters during muscle cell differentiation (Appendix). Numerous technologies derived from 3C are currently at the forefront of characterizing the three-dimensional organization of chromatin. We expect these and other methods to set the stage for a global understanding of chromatin in different cell types and disease states.

FUTURE DIRECTIONS

Based on the work in this Thesis, as well as other recent studies, a model for the three-dimensional organization of chromatin is emerging. In general, it appears that active chromatin marked by H3K4me1, H3K4me2, H3K4me3, H3K36me3, H4K20me1, and H3 and H4 acetylation is located in the center of the nucleus, and contains early replicating DNA (Ryba, Hiratani et al.). In contrast, inactive chromatin marked by H3K9me2 and H3K9me3 is located at the periphery of the nucleus, contacts the nuclear lamina, and contains late replicating DNA (Chapter 4) (Guelen, Pagie et al. 2008; Wen, Wu et al. 2009).

In addition, various specialized bodies seem to be located throughout the nuclear milieu. For instance, Polycomb bodies are thought to be discrete foci that contain Polycomb proteins, H3K27me3-marked chromatin, and silenced genes (Chapter 2; Chapter 4) (Hawkins, Hon et al. ; Sexton, Schober et al. 2007; Pauler, Sloane et al. 2009). Transcription factories are nuclear hotspots that contain 6 to 8 active polymerases, multiple active genes, and corresponding transcription factors (Schoenfelder, Sexton et al. ; Osborne, Chakalova et al. 2004; Misteli 2007). To further complicate the picture, there is evidence that each chromosome is located in a territory, the nucleolus has associated domains, and specific genomic regions can loop to come into contact (Appendix) (Nemeth, Conesa et al. ; Meaburn and Misteli 2007). It is intriguing to think that each cell type or disease state may have a unique three-dimensional chromatin organization that is particularly suited for its function.

Initial evidence already suggests that global chromatin structure reflects cell state and changes during cell differentiation. For instance, ES cells are characterized by globally open chromatin that is enriched in active marks and loosely associated with architectural proteins

(Meshorer and Misteli 2006). Polycomb in ES cells poises developmental genes for rapid activation or repression upon differentiation, and buffers transcriptional noise to maintain pluripotency (Chapter 2) (Bernstein, Mikkelsen et al. 2006; Chi and Bernstein 2009). Indeed, remodeling chromatin seems to be one of the crucial changes during reprogramming (Ang, Gaspar-Maia et al. 2011; Onder, Kara et al. 2012). Thus, it would be interesting to use DamID-seq to chart chromatin proteins in various cell types during differentiation. This would further test the model that cellular differentiation corresponds to global chromatin reorganization, and perhaps identify unique signatures for each cell type.

Furthermore, evidence suggests that chromatin structure reflects aberrant changes in diseased cells. For instance, Wilms tumor cells seem to have a similar chromatin landscape as renal stem cells, implying that they may originate from arrested development of these cells (Aiden, Rivera et al. 2010). Indeed, tumors in general seem to have gross changes in chromatin domains, as seen by histological changes evident within nuclei and the notable loss of LOCKs in cancer cell lines (Wen, Wu et al. 2009). Additionally, the three-dimensional structure of the genome affects the frequency of chromosomal translocations (Branco and Pombo 2006). For instance, two recent studies show that androgen-induced proximity of genomic loci may influence the frequency of aberrant fusion (Lin, Yang et al. 2009; Mani, Tomlins et al. 2009). Thus, it may be fruitful to use DamID-seq to map chromatin proteins in various disease states. As DamID-seq is able to chart mutant proteins by fusing the Dam enzyme to a mutated sequence of the protein, it may be particularly suited to study chromatin in disease models.

Currently, DamID-seq technology profiles a protein of interest in an ensemble of cells over three days. To detect rapid changes during differentiation or aberrant changes within a small cell population, greater sensitivity is needed. We note that for ChIP-seq, efforts are already

underway to minimize the amount of starting material needed and improve the sensitivity of assays (Adli, Zhu et al. ; Goren, Ozsolak et al.). Analogous improvements to DamID-seq may bring us closer to characterizing chromatin proteins in small cell numbers. Additionally, since the Dam protein dynamically marks its binding sites in a living cell, it may one day be possible to track this footprint in real time using imaging.

Advancements in DamID-seq for characterizing chromatin proteins, along with new technologies for examining three-dimensional chromatin structure, may eventually lead us to reach the ultimate goal: to characterize chromatin structure for every individual cell at any time point. I speculate that in the future, the field may assemble a catalog of the three-dimensional chromatin organization of most cell types in the human body. We may even be able to track changes in this structure as cells differentiate; for instance, we may visualize chromatin compacting and moving towards the nuclear periphery as a pluripotent cell becomes specialized. Then by relating this chromatin structure to genome function, we will be able to interpret the human genome more fully. In parallel, researchers may also characterize chromatin organization in various disease states. By comparing these maps to those of healthy cells, one may be able to pinpoint aberrant changes, and better understand, detect, and treat disease.

REFERENCES

Adli, M., J. Zhu, et al. "Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors." Nat Methods **7**(8): 615-618.

Aiden, A. P., M. N. Rivera, et al. (2010). "Wilms tumor chromatin profiles highlight stem cell properties and a renal developmental network." Cell Stem Cell **6**(6): 591-602.

Ang, Y. S., A. Gaspar-Maia, et al. (2011). "Stem cells and reprogramming: breaking the epigenetic barrier?" Trends Pharmacol Sci **32**(7): 394-401.

Bernstein, B. E., T. S. Mikkelsen, et al. (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." Cell **125**(2): 315-326.

Branco, M. R. and A. Pombo (2006). "Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations." PLoS Biol **4**(5): e138.

Chi, A. S. and B. E. Bernstein (2009). "Developmental biology. Pluripotent chromatin state." Science **323**(5911): 220-221.

Goren, A., F. Ozsolak, et al. "Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA." Nat Methods **7**(1): 47-49.

Guelen, L., L. Pagie, et al. (2008). "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions." Nature **453**(7197): 948-951.

Hawkins, R. D., G. C. Hon, et al. "Distinct epigenomic landscapes of pluripotent and lineage-committed human cells." Cell Stem Cell **6**(5): 479-491.

Lin, C., L. Yang, et al. (2009). "Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer." Cell **139**(6): 1069-1083.

Mani, R. S., S. A. Tomlins, et al. (2009). "Induced chromosomal proximity and gene fusions in prostate cancer." Science **326**(5957): 1230.

Meaburn, K. J. and T. Misteli (2007). "Cell biology: chromosome territories." Nature **445**(7126):

379-781.

Meshorer, E. and T. Misteli (2006). "Chromatin in pluripotent embryonic stem cells and differentiation." Nat Rev Mol Cell Biol **7**(7): 540-546.

Misteli, T. (2007). "Beyond the sequence: cellular organization of genome function." Cell **128**(4): 787-800.

Nemeth, A., A. Conesa, et al. "Initial genomics of the human nucleolus." PLoS Genet **6**(3): e1000889.

Olins, D. E. and A. L. Olins (2003). "Chromatin history: our view from the bridge." Nat Rev Mol Cell Biol **4**(10): 809-814.

Onder, T. T., N. Kara, et al. (2012). "Chromatin-modifying enzymes as modulators of reprogramming." Nature **483**(7391): 598-602.

Osborne, C. S., L. Chakalova, et al. (2004). "Active genes dynamically colocalize to shared sites of ongoing transcription." Nat Genet **36**(10): 1065-1071.

Pauler, F. M., M. A. Sloane, et al. (2009). "H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome." Genome Res **19**(2): 221-233.

Ryba, T., I. Hiratani, et al. "Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types." Genome Res **20**(6): 761-770.

Schoenfelder, S., T. Sexton, et al. "Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells." Nat Genet **42**(1): 53-61.

Sexton, T., H. Schober, et al. (2007). "Gene regulation through nuclear organization." Nat Struct Mol Biol **14**(11): 1049-1055.

Wen, B., H. Wu, et al. (2009). "Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells." Nat Genet **41**(2): 246-250.

Appendix:

Differentiation-Specific Looping Interactions between Distant Enhancers and Muscle Gene Promoters

Differentiation-Specific Looping Interactions between Distant Enhancers and Muscle Gene Promoters

This Chapter is revised from the following publication to reflect the contributions of Vicky W. Zhou:

"Distant cis-Regulatory Elements in Human Skeletal Muscle Differentiation." (2011) Genomics (98) 401-411

Rachel Patton McCord[1,5], Vicky W. Zhou[1,4], Tiffany Yuh[1,6], Martha L. Bulyk[1,2,3,5]*

1. Division of Genetics, Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, MA 02115, USA.

2. Department of Pathology, Brigham & Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

3. Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA

4. Biological and Biomedical Sciences Graduate Program, Harvard Medical School, Boston, MA 02115, USA

5. Harvard University Graduate Biophysics Program, Cambridge, MA 02138, USA

6. Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

RP McCord current affiliation: Program in Gene Function and Expression, University of Massachusetts Medical School, Worcester, MA 01605.

* To whom correspondence should be addressed:

Martha L. Bulyk, Ph.D.

Email: mlbulyk@receptor.med.harvard.edu

ABSTRACT

Identifying gene regulatory elements and their target genes in human cells remains a significant challenge. Despite increasing evidence of physical interactions between distant regulatory elements and gene promoters in mammalian cells, many studies consider only promoter-proximal regulatory regions. Cis-regulatory modules (CRMs) that are distant (>20 kb) from muscle gene promoters are common and are more likely than proximal promoter regions to show differentiation-specific changes in myogenic TF binding. We find that two of these distant CRMs, known to activate transcription in differentiating myoblasts, interact physically with gene promoters (PDLIM3 and ACTA1) during differentiation. Our results highlight the importance of considering distal CRMs in investigations of mammalian gene regulation and support the hypothesis that distant CRM-promoter looping contacts are a general mechanism of gene regulation.

INTRODUCTION

The identification of genomic sequence regions that regulate genes in a condition-specific manner is essential to understanding how the same genome sequence can give rise to the diversity of cell types and functions observed in an organism. In an organism with a small genome, such as yeast, the majority of gene regulation can be explained by transcription factor (TF) binding and chromatin modifications within approximately 600 bp to 1 kb of DNA sequence upstream of the regulated gene (Chua, Robinson et al. 2004; Zhu, Byers et al. 2009). In metazoans, numerous prior studies in a range of organisms from sea urchin to mammals have

identified cis-regulatory modules (CRMs), consisting of clusters of TF binding sites, located next to (or within the introns of) the genes whose expression they regulate (Davidson 2001). However, as compared to the yeast genome, metazoan genomes, in particular mammalian genomes, have a much higher proportion of noncoding sequence, and recent research has highlighted the importance of more distant CRMs in gene regulation within these genomes (Dostie, Richmond et al. 2006; Kumaran, Thakar et al. 2008; Sexton, Bantignies et al. 2009). Much detailed work on distantly located transcriptional enhancers in Drosophila has shown the importance of such distant regulatory elements and the ways that they can be directed to their target genes by insulator boundaries and promoter targeting sequences (Lin, Lin et al. 2010). In mammalian systems, the rules governing enhancer–promoter links are less evident, but certain examples of distant regulatory elements have been studied in detail. For example, the locus control region (LCR) of themurine β-globin locus forms a GATA-1-dependent looping interaction with actively transcribed globin gene promoters located approximately 40–60 kb away in erythroid cells (Tolhuis, Palstra et al. 2002; Vakoc, Letting et al. 2005). Such results suggest that a complete understanding of gene regulation will require searching for CRMs distant from target genes and further studies of how these CRMs are directed to and regulate their target genes.

The differentiation of human skeletal myoblasts into mature muscle fibers requires the coordinated regulation of many genes and is essential for the development and maintenance of proper muscle function. This differentiation process can be easily induced and monitored in cell culture, making it a tractable model system for investigating the regulatory mechanisms underlying dynamic, tissue specific gene expression in human. Prior studies have measured changes in gene expression and TF binding during differentiation in primary human skeletal

muscle cells or in the similar C2C12 mouse skeletal muscle cell line (Bergstrom, Penn et al. 2002; Blais, Tsikitis et al. 2005; Cao, Kumar et al. 2006; Warner, Philippakis et al. 2008; Cao, Yao et al. 2010). These studies have shown that a master regulatory TF, MyoD (encoded by MYOD1), initiates the cascade of gene regulation that leads to fusion of undifferentiated myoblast cells into multinucleated, elongated myotubes with developed contractile elements (Berkes and Tapscott 2005; Cao, Kumar et al. 2006). Other TFs in the myogenic regulatory factor (MRF) basic helix-loop-helix (bHLH) family – myogenin (MyoG, encoded by MYOG), Myf5 (MYF5), and MRF4 (MYF6) – some of which are activated directly by MyoD, work together with MyoD and other TFs such as Serum Response Factor (SRF) and the Myocyte Enhancer Factor 2 (Mef2) family to regulate the expression of genes involved in the continuation and completion of this differentiation process (Berkes and Tapscott 2005; Cao, Kumar et al. 2006; Gianakopoulos, Mehta et al. 2011). Using the sets of differentially expressed genes, conservation of sequence across species, and the DNA binding site motifs of myogenic TFs, CRMs responsible for coordinating changes in expression during myogenic differentiation have been predicted computationally (Thompson, Palumbo et al. 2004; Sun, Chen et al. 2006; Warner, Philippakis et al. 2008). However, unlike the β-globin locus in which some of the classic distant regulatory interactions have been characterized previously, little is known about the role of distant regulatory modules in human skeletal muscle differentiation.

Previous efforts to characterize the transcriptional regulatory network in skeletal muscle differentiation have measured binding of myogenic TFs using chromatin immunoprecipitation followed by microarray hybridization or sequencing (ChIP-chip or ChIP-Seq, respectively) in murine C2C12 cell lines (Blais, Tsikitis et al. 2005; Cao, Kumar et al. 2006; Cao, Yao et al. 2010) and have tested the regulatory function of some predicted TF-bound elements with

reporter assays (Sun, Chen et al. 2006). While the results of these studies have revealed previously unknown regulatory connections between TFs and differentiation processes, they tended to focus on TF binding near promoters (approximately 1–4 kb upstream of transcriptional start sites (TSSs)) (Blais, Tsikitis et al. 2005; Cao, Kumar et al. 2006; Sun, Chen et al. 2006). In contrast, methods to predict CRMs involved in myogenic differentiation have identified candidate regulatory modules by searching sequences up to 10 kb (Thompson, Palumbo et al. 2004) or 50 kb (Warner, Philippakis et al. 2008) away from TSSs. A handful of the more distantly located predicted CRMs (approximately 20–30 kb upstream or downstream of TSSs) were found to activate expression in reporter assays specifically during myogenic differentiation and to be bound by myogenic TFs (Warner, Philippakis et al. 2008). These results suggest that gene regulatory myogenic TF binding occurs at locations distant from the proximal promoter regions that have been the focus of most prior studies.

Here, we determined whether distant CRMs with no microRNA or protein-coding gene promoters nearby can form long-range looping interactions with muscle gene promoters. We examined two distant CRMs previously found to drive gene expression specifically during myogenic differentiation as case examples. We found that these CRMs form differentiation-specific physical interactions with their closest target genes. Our results provide further support that a complete understanding of transcriptional regulation in mammals will require consideration of CRMs located distant from their target genes in the genome sequence.

RESULTS

Differentiation-specific looping interactions between distant CRMs and muscle gene promoters

We used Chromosome Conformation Capture (3C) (Dekker, Rippe et al. 2002; Hagege, Klous et al. 2007) experiments to investigate whether the putative myogenic CRMs that we found located distant from muscle genes form physical interactions with their adjacent, putative target genes during myogenic differentiation. We selected two CRMs that were bound by MyoG at 48 h after induction of differentiation, exhibited H3K4me1 and H3K27Ac enhancer-associated histone marks, were found to drive expression of a reporter gene during differentiation (Warner, Philippakis et al. 2008), and were located at least 20 kb away from the TSS of the closest known protein-coding gene. The "ACTA1 CRM" (Figure A.1A) is located 23 kb downstream of the TSS of ACTA1, a gene which encodes the skeletal muscle alpha actin protein, an essential part of the contractile element in muscle. The "PDLIM3/SORBS2 CRM" (Figure A.1B) is located 32 kb upstream of PDLIM3, which encodes a cytoskeletal organizing protein localized to the Z line of the sarcomere and which may also regulate the myogenic differentiation transcriptional network by affecting the nuclear localization of SRF (Pomies, Pashmforoush et al. 2007). This CRM is also located 240 kb downstream of SORBS2, another gene differentially expressed during myogenic differentiation.

To assay the physical interactions between these two CRMs and their adjacent genes, we performed 3C experiments on human muscle cells 48 h before, immediately ("0 h") before, and 48 h after induction of differentiation by serum removal.

Figure A.1.

3C evidence that CRMs physically interact with muscle gene promoters 20–35 kb away during (0 h and 48 h) but not before (−48 h) differentiation. The y-axis shows the normalized ratio between the average quantity of the interaction-specific ligation PCR product vs. the BAC control. To allow relative comparisons between experiments, this ratio is normalized such that the interaction frequency between the fixed primer (located at 0) and its neighboring fragment is 1. Error bars=1 s.e.m. Restriction fragment boundaries are indicated above the graph. Note: not all restriction fragments are represented because some could not be assayed by good primers. (A) A 3C fixed primer at the ACTA1 promoter shows a significant interaction with a downstream CRM after, but not before, differentiation ($p<0.05$ by Student's t-test). (B) A 3C fixed primer at the CRM upstream of PDLIM3 shows a specific interaction with the PDLIM3 promoter after, but not before, differentiation ($p<0.05$ by Student's t-test). In both cases, the interaction is first established as the cells begin to differentiate (0 h) and then increases during differentiation (48 h).

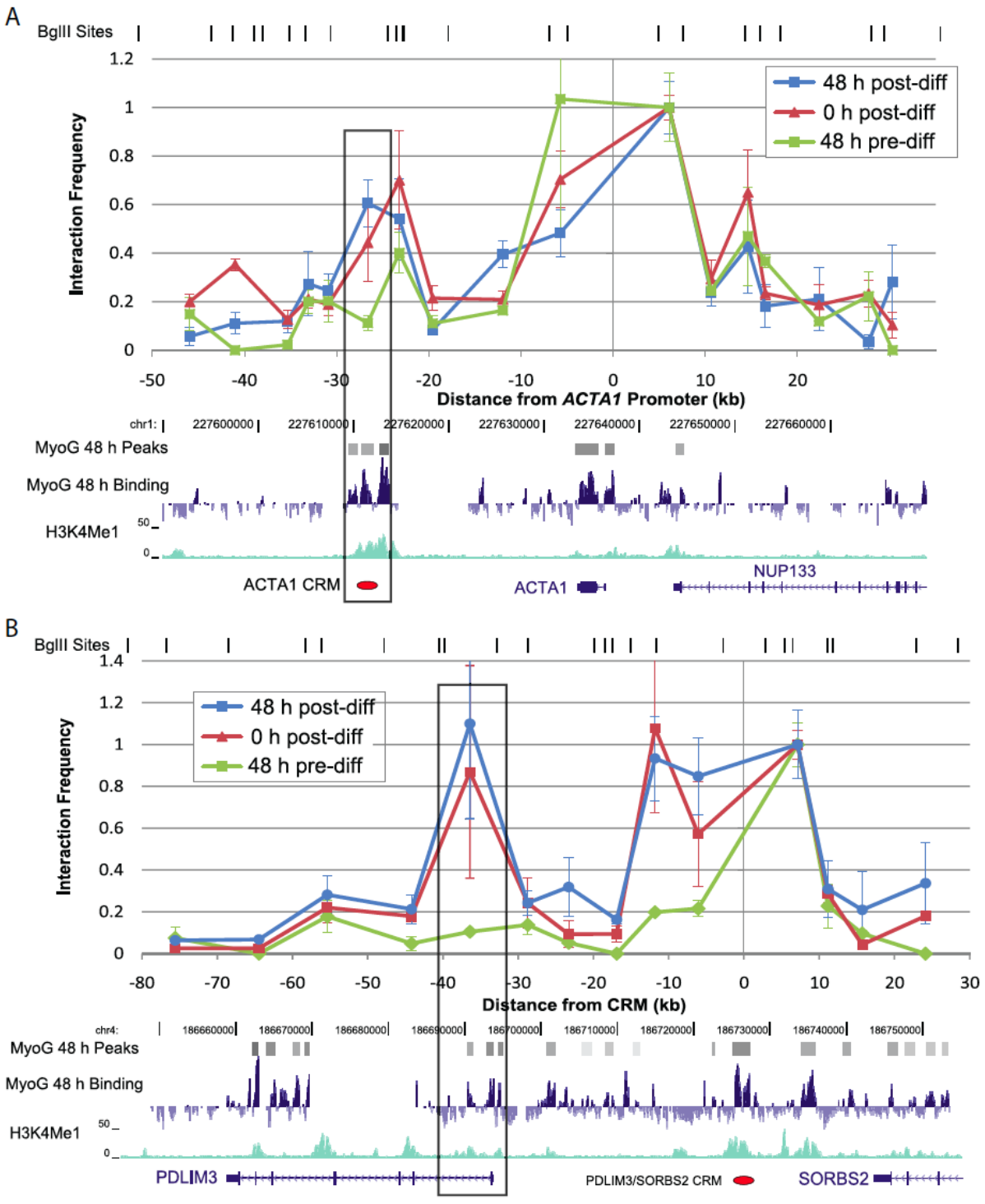

Figure A.1 (Continued).

We identified "differentiation-specific interactions" by searching for any peaks of interaction in the 0 h or 48 h post-differentiation cells, outside of the neighboring +/−15 kb that may interact highly by random collisions with the fixed bait, that were higher than the 48 h pre-differentiation interaction level (see Materials and Methods). The results of these experiments show evidence for a differentiation-specific interaction between the "ACTA1 CRM" and the proximal promoter region spanning from 1.6 kb upstream to 1.3 kb downstream of the ACTA1 TSS (Figure A.1A). We confirmed this looping interaction in cells at 48 h post-differentiation using two different restriction enzymes, BglII and AflII (Figure A.2). We also found a differentiation-specific interaction between the "PDLIM3/SORBS2 CRM" and the region spanning from 3 kb upstream to 4 kb downstream of the PDLIM3 TSS (Figure A.1B). This interaction was observed in primary myoblasts obtained from two different individuals.

The physical interactions between both of these previously identified CRMs (Warner, Philippakis et al. 2008) and their corresponding target gene promoters are absent 48 h before differentiation, but then become evident at 0 h, when the cells are confluent and elongating and differentiation is induced by serum removal. These CRM–promoter interaction peaks become more pronounced at 48 h after induction of differentiation (Figure A.1). These results suggest that a differentiation-induced chromosome conformation is established at these gene loci as the cells are approaching confluence and may be initiating the differentiation process even before stimulation of differentiation by serum removal. These physical interaction data parallel the previously published observation that the expression of the ACTA1 and PDLIM3 genes begins to increase from −48 h to 0 h (Warner, Philippakis et al. 2008) and our observations described above that predicted CRMs are often already bound by myogenic TFs at 0 h.

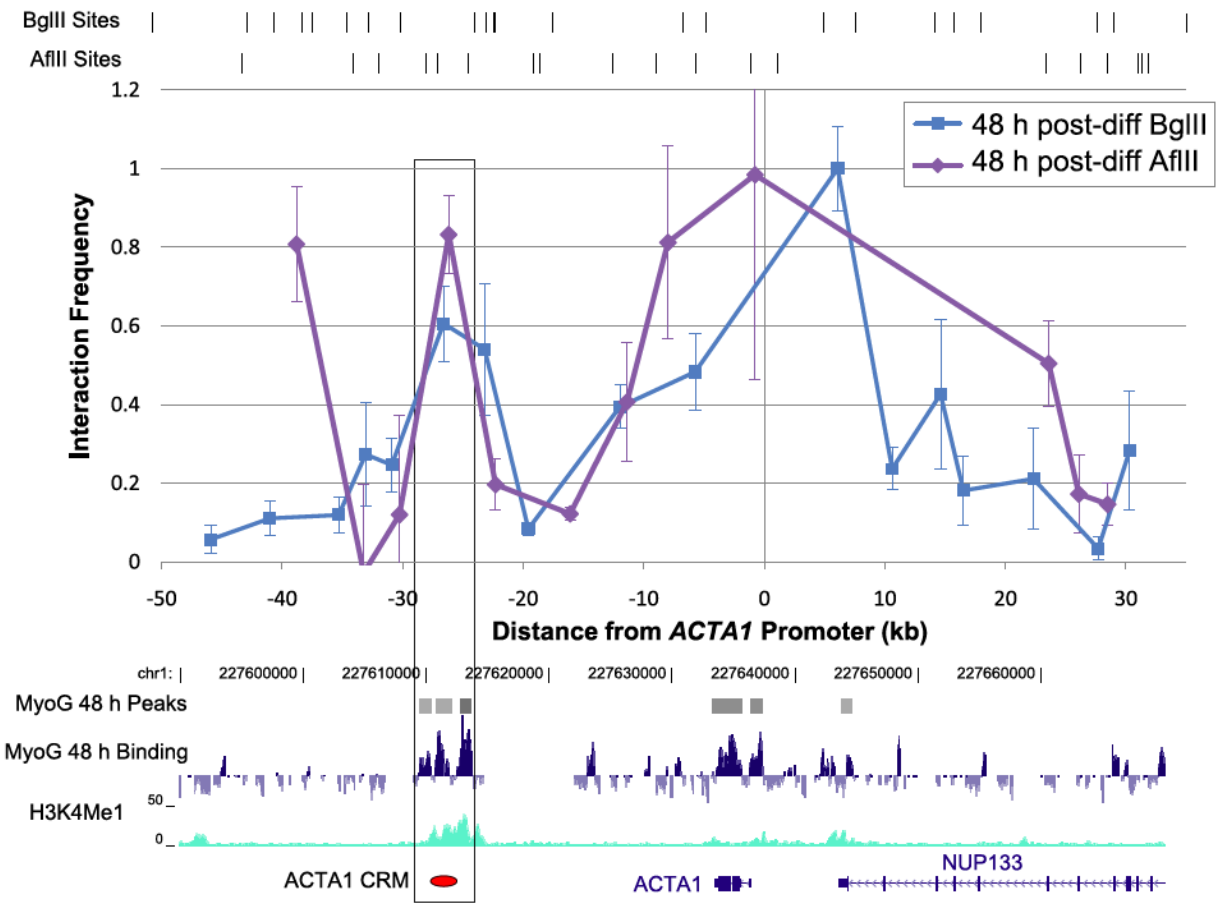

Figure A.2.

The interaction at 48 h post-differentiation between the *ACTA1* promoter and the CRM 25 kb downstream is confirmed by independent experiments with two restriction enzymes, AflII and BglII. AflII restriction sites are shown and narrow the region of the ACTA1 CRM interaction as compared to the BglII sites.

Interestingly, the PDLIM3/SORBS2 CRM was observed to interact only with the PDLIM3 promoter region 23 kb downstream of the CRM but not with either of the two alternative promoters of SORBS2 located 240 kb and 385 kb upstream of the PDLIM3/SORBS2 CRM. Thus, while the PDLIM3/SORBS2 CRM acts on its target promoter over a distance, the CRM selectively interacts with the closer promoter rather than the farther promoter. These observations suggest a model in which distant enhancers participate in activating gene expression by looping to contact the promoters of the genes closest to them. Alternatively, the promoter selectivity could be due to promoter sequence elements that differ between genes to specify the appropriate interaction or cell-type specific insulator architecture (Dean 2011).

DISCUSSION

In this study we found a variety of evidence that CRMs involved in myogenic differentiation can be located distant from the proximal promoter regions of muscle genes. Here, we found that two of these predicted and validated CRMs, the ACTA1 CRM and the PDLIM3/SORBS2 CRM, both located more than 20 kb from muscle genes, interact with adjacent gene promoters as assayed by 3C.

After distant CRMs are identified, a remaining challenge is to determine which genes they regulate and the mechanisms by which they affect gene expression. The differentiation-specific interactions observed in this study by 3C for the CRMs located 20–35 kb away from ACTA1 and PDLIM3 suggest a model in which these distal enhancers assist in activating gene expression by looping to contact the promoters of the closest genes. These are the first such distal interactions between a CRM and promoter that have been identified in mammalian muscle

differentiation. Combined with the mounting evidence of CRM–promoter interactions in other cell types and systems, these data suggest that long-distance physical interaction is a general mechanism of enhancer action rather than a mechanism specific to a particular system. That these enhancer–promoter interactions begin to form at 0 h of differentiation suggests that a differentiation-specific set of looping interactions may be established before the resulting changes in gene expression and cell phenotype begin. This is consistent with findings in other systems: in mouse erythroid progenitors, the β-globin LCR region exhibits interactions that are established before changes in gene expression occur (Phillips and Corces 2009), and interactions between enhancers and promoters implicated in hormone responsive gene expression in mouse adenocarcinoma cells are already present at lower levels before hormone treatment (Hakim, Sung et al. 2011).

When identifying target genes for CRMs outside proximal promoter regions, some previous studies have used the simplifying assumption that CRMs are likely to interact with gene promoters lying within a domain bounded by CTCF sites, on the basis that CTCF often acts as an insulator across which boundary interactions are less likely to occur (Heintzman and Ren 2009; Cao, Yao et al. 2010). This assumption matches experimental results in the case of the PDLIM3/SORBS2 CRM, which interacts with the PDLIM3 promoter in the same CTCF domain (as measured by a previous study in human skeletal muscle myoblasts (Ernst, Kheradpour et al. 2011)), but does not interact with the SORBS2 promoters that are separated from the CRM by several CTCF sites. However, the interaction we observed between the ACTA1 CRM and the ACTA1 promoter crosses a CTCF site, as does the interaction between the PDLIM3/SORBS2 CRM and the predicted CRM located 27 kb upstream of a SORBS2 promoter. Thus, our results

suggest that the links between CRMs and their target genes cannot always be predicted accurately by CTCF sites.

How enhancer–promoter interactions are established and why they are advantageous to myogenic differentiation are yet to be determined. It is possible that the myogenic TFs observed to bind to these interacting regions are also responsible for mediating these interactions. However, a full understanding of the interaction mechanism will require future work to search for sequence motifs recognized by other factors potentially involved in establishing chromosomal domains (CTCF, for example (Phillips and Corces 2009)), to experimentally purify factors associated with the interacting genomic regions (Dejardin and Kingston 2009), and to disrupt implicated DNA binding sites to test whether they are required for the physical interactions.

This study demonstrates that CRMs for myogenic differentiation distant from muscle gene promoters can physically interact with their target genes. These results highlight the importance of looking beyond the proximal promoter to understand the transcriptional regulation of genes involved in differentiation. These distant CRM–promoter interactions may relate to the global changes in genome organization observed as cells undergo differentiation (Rajapakse, Perlman et al. 2009). In the future, combining computational enhancer predictions, experimental tests of distant CRMs, and global measurements of chromosome conformation (Lieberman-Aiden, van Berkum et al. 2009) will help to further elucidate the mechanisms of gene regulation during this differentiation process.

MATERIALS AND METHODS

Proliferation and differentiation of human myoblasts in cell culture

Adult primary human skeletal myoblasts (Lonza) were grown in SkGM-2 medium (Lonza). Myogenic differentiation was stimulated by switching the culture medium to DMEM-F12 with 2% horse serum (Sigma) when the cells reached about 70% confluence. All time points referred to in this study are with respect to the time of switching to differentiation medium. The majority of the results presented here were obtained using adult male skeletal myoblasts; female skeletal myoblasts were used in one 3C experiment to confirm the reproducibility, and lack of gender-specificity, of the differentiation-specific physical interaction of the PDLIM3/SORBS2 CRM.

Chromosome Conformation Capture (3C)

3C experiments were performed following previously described protocols (Dostie, Richmond et al. 2006). Briefly, approximately 5×107 human muscle cells were crosslinked in a final concentration of 1% formaldehyde and harvested at each of two timepoints: 48 h before and 48 h after differentiation by serum deprivation. Chromatin extracted from these cells was digested with either an AflII or BglII restriction enzyme (New England Biolabs) and then incubated with T4 DNA ligase (New England Biolabs), such that regions of chromosome contact were ligated together. After reversing the formaldehyde crosslinks, physical interactions between genomic regions were detected with PCR primers specific for each ligated interaction product.

Primer pairs with melting temperatures of 56–60 °C, within 2 °C of each other, and of 35–50% GC content, were designed to uniquely amplify an approximately 100–300 bp region straddling the location of potential ligation between two restriction fragments. PCR-amplified products were detected by quantitative real-time PCR (qPCR) on an iCycler (BioRad) with SybrGreen Supermix for iQ (BioRad) and standard protocols. The quantification by qPCR was validated by visualization on an agarose gel and quantification using QuantityOne gel image quantification software (BioRad). The qPCR quantification results were utilized for the final analysis.

To control for differences in primer efficiency, DNA fragments generated from bacterial artificial chromosomes (BACs) spanning the genomic regions of interest were digested and randomly ligated (ACTA1: RP11-1111E20, PDLIM3/SORBS2: CTD-2559A19, CTD-2194A4, and RP11-78H20). Quantified PCR products from each 3C reaction were then normalized by the quantified results from the same PCR amplifications performed on this BAC control library. For those regions for which an ANOVA test of interaction frequencies within a given genomic region (excluding fragments within 15 kb of the fixed primer, which are likely to show high levels of interactions from random collisions (Dekker 2006)) indicated the presence of a significant peak ($p<0.05$) within the data, we tested the observed peaks for statistically significant differences in interaction level between timepoints using a Student's t-test.

ACKNOWLEDGEMENTS

The authors thank Job Dekker and Ye Zhan for providing training in the 3C technique, Jason Warner for assistance with array design and training in ChIP-chip experiments, Cherelle Walls for assistance with cell culture and ChIP-chip experiments, and Anthony Philippakis for

assistance with PhylCRM and array design. This work was supported by NIH/NHGRI grant # R21 HG005149 (M.L.B.). R.P.M. was supported in part by a National Science Foundation Graduate Research Fellowship. V.Z. was supported by a National Defense Science and Engineering Graduate Fellowship and a National Science Foundation Graduate Research Fellowship.

REFERENCES

Bergstrom, D. A., B. H. Penn, et al. (2002). "Promoter-specific regulation of MyoD binding and signal transduction cooperate to pattern gene expression." Mol Cell **9**(3): 587-600.

Berkes, C. A. and S. J. Tapscott (2005). "MyoD and the transcriptional control of myogenesis." Semin Cell Dev Biol **16**(4-5): 585-595.

Blais, A., M. Tsikitis, et al. (2005). "An initial blueprint for myogenic differentiation." Genes Dev **19**(5): 553-569.

Cao, Y., R. M. Kumar, et al. (2006). "Global and gene-specific analyses show distinct roles for Myod and Myog at a common set of promoters." EMBO J **25**(3): 502-511.

Cao, Y., Z. Yao, et al. (2010). "Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming." Dev Cell **18**(4): 662-674.

Chua, G., M. D. Robinson, et al. (2004). "Transcriptional networks: reverse-engineering gene regulation on a global scale." Curr Opin Microbiol **7**(6): 638-646.

Davidson, E. H. (2001). "Genomic Regulatory Systems: Development and Evolution." Academic Press, San Diego.

Dean, A. (2011). "In the loop: long range chromatin interactions and gene regulation, Brief." Funct. Genomics **10**: 3-10.

Dejardin, J. and R. E. Kingston (2009). "Purification of proteins associated with specific genomic Loci." Cell **136**(1): 175-186.

Dekker, J. (2006). "The three 'C' s of chromosome conformation capture: controls, controls, controls." Nat Methods **3**(1): 17-21.

Dekker, J., K. Rippe, et al. (2002). "Capturing chromosome conformation." Science **295**(5558): 1306-1311.

Dostie, J., T. A. Richmond, et al. (2006). "Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements." Genome Res **16**(10): 1299-1309.

Ernst, J., P. Kheradpour, et al. (2011). "Mapping and analysis of chromatin state dynamics in nine human cell types." Nature **473**(7345): 43-49.

Gianakopoulos, P. J., V. Mehta, et al. (2011). "MyoD directly up-regulates premyogenic mesoderm factors during induction of skeletal myogenesis in stem cells." J Biol Chem **286**(4): 2517-2525.

Hagege, H., P. Klous, et al. (2007). "Quantitative analysis of chromosome conformation capture assays (3C-qPCR)." Nat Protoc **2**(7): 1722-1733.

Hakim, O., M. H. Sung, et al. (2011). "Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements." Genome Res **21**(5): 697-706.

Heintzman, N. D. and B. Ren (2009). "Finding distal regulatory elements in the human genome." Curr Opin Genet Dev **19**(6): 541-549.

Kumaran, R. I., R. Thakar, et al. (2008). "Chromatin dynamics and gene positioning." Cell **132**(6): 929-934.

Lieberman-Aiden, E., N. L. van Berkum, et al. (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." Science **326**(5950): 289-293.

Lin, Q., L. Lin, et al. (2010). "Chromatin insulator and the promoter targeting sequence modulate the timing of long-range enhancer-promoter interactions in the Drosophila embryo." Dev Biol **339**(2): 329-337.

Phillips, J. E. and V. G. Corces (2009). "CTCF: master weaver of the genome." Cell **137**(7): 1194-1211.

Pomies, P., M. Pashmforoush, et al. (2007). "The cytoskeleton-associated PDZ-LIM protein, ALP, acts on serum response factor activity to regulate muscle differentiation." Mol Biol Cell **18**(5): 1723-1733.

Rajapakse, I., M. D. Perlman, et al. (2009). "The emergence of lineage-specific chromosomal topologies from coordinate gene regulation." Proc Natl Acad Sci U S A **106**(16): 6679-6684.

Sexton, T., F. Bantignies, et al. (2009). "Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation." Semin Cell Dev Biol **20**(7): 849-855.

Sun, Q., G. Chen, et al. (2006). "Defining the mammalian CArGome." Genome Res **16**(2): 197-207.

Thompson, W., M. J. Palumbo, et al. (2004). "Decoding human regulatory circuits." Genome Res **14**(10A): 1967-1974.

Tolhuis, B., R. J. Palstra, et al. (2002). "Looping and interaction between hypersensitive sites in the active beta-globin locus." Mol Cell **10**(6): 1453-1465.

Vakoc, C. R., D. L. Letting, et al. (2005). "Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1." Mol Cell **17**(3): 453-462.

Warner, J. B., A. A. Philippakis, et al. (2008). "Systematic identification of mammalian regulatory motifs' target genes and functions." Nat Methods **5**(4): 347-353.

Zhu, C., K. J. Byers, et al. (2009). "High-resolution DNA-binding specificity analysis of yeast transcription factors." Genome Res **19**(4): 556-566.